Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 3, No. 1, 2023

Prediction of Telecom Customer Churn Based on MIPCA-XGBoost Method

Chen Zhuo*

School of Control and Computer Engineering, North China Electric Power University, Beijing 102200, China
* **Corresponding author Email:** cherrybombczz@163.com

Abstract: In order to solve the problem that the nonlinear information of data in the field of telecom customer churn prediction is not fully used, or even ignored, which leads to inaccurate prediction, this paper introduces the mutual information feature selection method (MIPCA) to filter the features and reduce the dimensions of customer data, and proposes an XGBoost method based on the mutual information feature selection method(MIPCA-XGBoost), which improves the accuracy of the prediction results. By using the data set of telecom industry customers published on Kaggle website, compares the prediction result of this method with that of machine learning algorithms commonly used in this field, and proves the accuracy, recall and F_Score of MIPCA-XGBoost method is higher than other algorithms.

Keywords: Customer churn; Telecom customers; Mutual information feature selection; XGBoost algorithm.

1. Introduction

Nowadays, the market of China's telecom industry is becoming increasingly saturated, the competition between enterprises is becoming more and more fierce, and the customer churn rate is gradually rising. Many enterprises focus on how to use novel marketing models to attract new customers and further develop them into loyal customers. However, studies show that the cost of time and money required for a company to develop a new customer is far greater than the cost of maintaining an existing customer [1]. Therefore, it is imperative for enterprises to build a model that can accurately predict the customer churn tendency, understand the customer churn tendency in time and adjust the customer maintenance strategy to reduce the customer churn rate.

Customer churn is a complex problem, and the prediction methods for different industries are different. At present, there are a large number of domestic literatures on customer churn, involving a wide range of fields. Zhang Lili [2] and others used decision tree algorithm to predict airline customer churn, and successfully improved the seating rate; Yan Chun [3] and others used BP-Adaboost algorithm to predict the clustered life insurance industry customers, providing a higher prediction accuracy; In the field of e-commerce, Wu Yongchun [4] fused multiple methods to establish a prediction model, which shortened the prediction time. However, the customer data in the telecommunications industry has the characteristics of large quantity and high dimension. Although the above research has made important contributions to the research on customer churn prediction in the fields of aviation, life insurance, e-commerce and so on, it does not mine the important features of the data set, resulting in information redundancy and even dimension disaster [5], which will have an impact on the customer churn prediction in the telecommunications industry and reduce its prediction efficiency and accuracy.

At present, the data feature selection in the field of customer churn prediction is generally based on the traditional statistical principal component analysis and linear discriminant method [6]. Its advantages are that the theoretical basis is solid and the method is simple and easy to operate, but the nonlinear relationship between attributes is ignored, the information is not fully utilized, and even important information is lost. The mutual information feature selection method can retain most of the original information while effectively reducing the dimension, and take into account the nonlinear relationship between variables [7]. In addition, XGBoost algorithm has full application in short-term photovoltaic power generation prediction [8-11], flight delay prediction [12-14], food safety risk prediction [15-18] and other prediction fields. These studies show that XGBoost has the advantages of high flexibility, fast execution speed, and high accuracy of prediction results. Based on this, this paper proposes an XGBoost prediction method based on mutual information feature selection to solve the problem of customer churn in the telecommunications industry, and compares the accuracy, recall, and F_ Score and other indicators, and the analysis proves the effectiveness of the method proposed in this paper in telecom customer churn prediction.

2. MIPCA-XGBoost customer churn prediction method

2.1. MIPCA method

In information theory, mutual information is a measure of interdependence between random variables, which can also be understood as the amount of information of another random variable contained in one random variable [11]. In feature selection, the principal component analysis method in traditional statistics is to analyze the linear relationship between two variables, but cannot reflect the nonlinear relationship in the data. Therefore, this paper considers introducing mutual information to feature selection, which is very helpful for evaluating the interdependence between variables, and is no longer limited to the linear relationship. Here we introduce information entropy H, which is used to represent the uncertainty of a random variable.

Suppose there are random variables X and Y in data set M, and there are joint distributions $p(x,y)$ and marginal distributions $p(x)$ and $p(y)$, then there are

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \quad (1)$$

Where, $H(X|Y)$and $H(Y|X)$ are conditional information entropy, which represents the information entropy of its own information after deducting other conditions, which can be expressed as

$$H(X|Y) = -\sum_{x\in X} p(x,y) log p(x|y) \quad (2)$$

$$H(Y|X) = -\sum_{y\in Y} p(x,y) log p(y|x) \quad (3)$$

For random variables X and Y, the amount of information of another variable contained in one variable can be expressed by mutual information [11], which is defined as

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (4)$$

The information entropy of random variables X and Y can be expressed as

$$H(X) = -\sum_{x\in X} p(x) log p(x) \quad (5)$$

$$H(Y) = -\sum_{y\in Y} p(y) log p(y) \quad (6)$$

Therefore, we can further obtain the mutual information $I(X,Y)$ of random variables X and Y as

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \quad (7)$$

In the principal component analysis method, the mutual information matrix is used to replace the covariance matrix, and the relationship between the eigenvector and the eigenvalues of the mutual information feature selection method is obtained as

$$A^T \Sigma_I A = B \quad (8)$$

Among them, Σ_I is the mutual information matrix corresponding to the data set M, A is the matrix formed when the eigenvector of the mutual information matrix is a column vector, and B is the eigenvalue matrix of the mutual information matrix. In Σ_I, the elements on the diagonal represent the self-information of variables, while the elements on the non-diagonal represent the mutual information of variables. When the two variables are not related, the mutual information is 0. The principal component c of mutual information matrix Σ_I is

$$c_k = \alpha_k^T x \quad (9)$$

Among them, $\alpha_k^T (k = 1,2,\dots,n) \in A^T$, is conversion coefficient of c_k. Define the contribution rate of the kTH feature as σ_k. The formula

$$\sigma_k = \frac{\mu_k}{\sum_{k=1}^{n} \mu_k} \quad (10)$$

μ_k is the k TH largest eigenvalue of the mutual information matrix. In this paper, the first m principal components whose cumulative contribution rate of β is about 90% are selected as the output of the feature selection results, and the dimensionality reduction of the data set is realized.

2.2. Principles of XGBoost model

XGBoost algorithm is a model that uses CART tree as the base learner for training and combines multiple base learners to construct a strong classifier. XGBoost is used to train the data set, divide the sample data into each leaf node according to different classification characteristics, calculate the gain value of the tree model before and after classification, and finally obtain a training model with the minimum loss value. The objective function can be expressed as

$$Obj = \sum_{i=1}^{n} l(x, \hat{x_i}) + \sum_{t=1}^{T} \Omega(f_t) \quad (11)$$

Where l is the loss function, x_i is the true value of the i-th sample data, $\hat{x_i}$ is the predicted value of the ith sample, n is the total number of samples, $\Omega(f_t)$ is the penalty term controlling the complexity of the model in the t th tree, and T is the number of training trees.

XGBoost is a superposition training model. When training the t-th tree, the objective function of the t-th tree is

$$Obj_t = \sum_{i=1}^{n} l\left(x_i, \hat{x_i}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + c \quad (12)$$

$$f_t(x_i) = w_{q(x_i)} \quad (13)$$

Where, $f_t(x_i)$ is the weight of sample x_i in the t-th tree, and its formula is shown in Equation (13), c is a constant, $q(x_i)$ is the position of the ith sample on the t-th tree, which is specifically represented by the number of leaf nodes it falls on, and $w_{q(x_i)}$ represents the weight of the leaf node.

The second-order Taylor series is used to expand the loss function, and the objective function after processing is

$$Obj_t = \sum_{k=1}^{K} \left[G_k w_k + \frac{1}{2}(H_k + \lambda) w_k^2\right] + \gamma K \quad (14)$$

$$\sum_{i\in I_k} g_i = G_k \quad (15)$$

$$\sum_{i\in I_k} h_i = H_k \quad (16)$$

Where, I_k represents all samples falling on the kTH leaf node, and g_i and h_i are the replacements of the first and second derivatives of the loss function respectively. For each sample that has been trained t-1 times, its g and h are known. G_k andH_k are the sum of g and h of all samples on k leaf nodes respectively.

By observing the simplified objective function formula, it can be seen that this is a quadratic equation with one variable, and the extreme value is

$$w^* = -\frac{G_k}{H_k + \lambda} \quad (17)$$

By substituting the extreme value into equation (14), the expression of leaf nodes in the tree to be trained can be written as

$$Obj = -\frac{1}{2}\sum_{k=1}^{K} \frac{G_k^2}{H_k + \lambda} + \gamma K \quad (18)$$

XGBoost model calculates the Gain value by calculating $Obj_{OLD} - Obj_{NEW}$:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \quad (19)$$

The higher the gain score, the higher the feature importance score of the split node, and the more important its corresponding feature.

2.3. MIPCA-XGBoost step

The procedure is as follows:

1)Feature screening was performed on the data set by MIPCA method, and the first m principal components were obtained as new features, and the dimensionality reduction of the initial data set was realized;

2)Randomly divide the new data set after feature selection processing, and the divided training set is used as the data input of the model, and the test set is used to verify the trained model. The flow chart of MIPCA-XGBoost is shown in Figure 1.

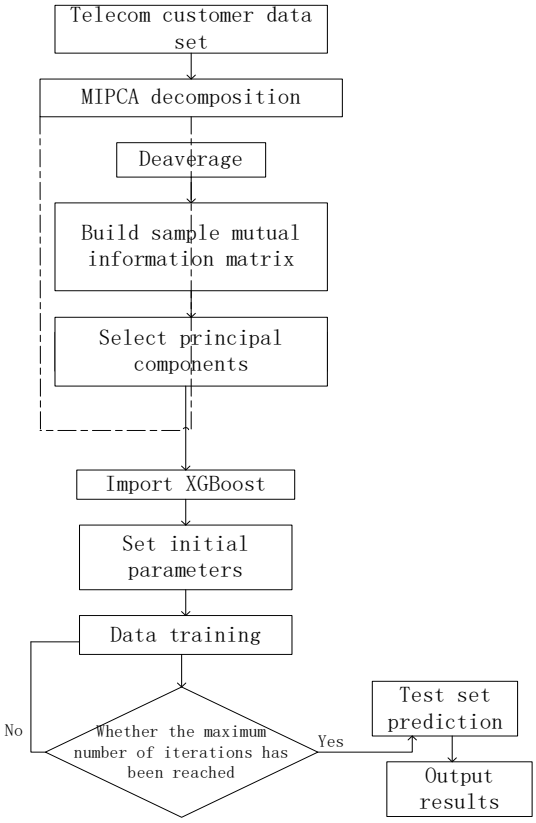

Fig. 1 The flow chart of MIPCA-XGBoost method

3. Data preparation

3.1. Experimental data

The telecom industry customer data set used in this paper has 48,712 pieces of data, including 9876 pieces of lost data and 38,836 pieces of non-lost data. The amount of lost data is about 20% of the total data set. In the experiment, the data set is randomly divided into 40% test set and 60% training set, which comes from Kaggle website. Each piece of data in the data set includes 57 characteristics and attributes, such as account opening time, remaining accounts, recharge period, online duration, and total call duration. Some attributes are shown in Table 1.

Table 1. Part properties of dataset

Data classification	Data attributes
Basic Information	Customer ID, age, package, account opening time
Billing information	Account balance, recharge period, credit limit, etc
Network duration	Time spent online, time spent online on weekdays, time spent online on weekends, etc
Communication information	Number of calls, total call duration, number of calls, number of calls, etc
Network Usage information	Uplink traffic, downlink traffic, traffic overflow, etc
Status of complaints	The number of complaints, the response time of complaints, etc

The data is preprocessed, 0 is used to fill the missing values in the data attributes, unique thermal coding is adopted for discrete attributes, and standardization is carried out for continuous attributes, so that the processed data can meet the standard normal distribution.

3.2. Feature selection

MIPCA method was used to select the feature of the data set, and the feature selection results of the data set were compared with the principal component analysis method. The comparison results were expressed as the principal component characteristic value μ and the cumulative contribution rate β, as shown in Table 2.

Table 2. Principal component eigenvalue and cumulative contribution rate

Principal component	PCA		MIPCA	
	μ	β	μ	β
PC1	16.34	30.07%	22.23	43.30%
PC2	8.56	46.87%	8.34	56.67%
PC3	5.23	57.13%	4.27	64.71%
PC4	4.12	65.22%	3.68	70.35%
PC5	2.85	69.81%	2.64	76.75%
PC6	2.21	73.15%	1.93	80.65%
PC7	1.80	76.68%	1.43	83.77%
PC8	1.48	79.59%	1.04	86.34%
PC9	1.23	82.64%	0.84	88.76%
PC10	1.01	84.76%	0.65	89.93%
PC11	0.88	85.60%	0.63	91.14%
PC12	0.76	85.72%	0.57	92.22%
PC13	0.64	86.43%	0.54	93.27%
PC14	0.64	87.14%	0.49	94.12%
PC15	0.62	87.80%	0.44	95.03%
PC16	0.58	88.47%	0.37	95.87%
PC17	0.56	89.22%	0.34	96.21%
PC18	0.55	89.75%	0.32	96.54%
PC19	0.55	90.28%	0.29	96.84%

From the extraction results of principal components in Table 2, it can be seen that MIPCA has a higher cumulative contribution rate when principal components of the same dimension are selected. For example, to achieve a cumulative contribution rate of 90%, PCA method needs to select 19-dimensional principal components, while MIPCA method reduces the value to 11-dimensional, which indicates that MIPCA method can make full use of nonlinear information. Thus, the original information can be retained to a greater extent, which is helpful to improve the accuracy of subsequent prediction. In this paper, when the cumulative contribution rate reaches 90%, the first 11 dimensional principal components screened by MIPCA method are selected for XGBoost customer loss prediction.

4. Customer churn forecast

4.1. Evaluation index

According to the confusion matrix, the prediction accuracy E_{acc}, the recall rate E_{recall}, and the F_Score were calculated as the evaluation index of the algorithm. The calculation formula of each evaluation index is shown as follows, and the classification of customer state prediction results [12] is shown in Table 3.

Table 3. Customer status classification

True customer status	Loss of prediction results	The prediction result is not churn
Actual loss	TP	FN
Actual non-churn	FP	TN

$$E_{acc} = \frac{TP+TN}{TP+TN+FP+FN} \quad (20)$$

$$E_{recall} = \frac{TP}{TP+FN} \quad (21)$$

$$F\_Score = \frac{2\times TP}{2\times TP+FP+FN} \quad (22)$$

4.2. Important parameter Settings of XGBoost algorithm

The experiment running environment in this paper is a 64-bit Windows10 operating system, and the specific hardware Settings are as follows: 16 GB memory and 12th Gen Intel(R) Core(TM) i5-12490F @ 3.00 GHz CPU. python3 and related toolkits were used in the experiments. After many experiments, the parameters of XGBoost are finally set in Table IV.

Table 4. The important parameters of XGBoost

Parameter	Parameter configuration
max-depth	5
min-child-weight	3
learning-rate	0.3
gamma	0.1
subsample	0.76
n-estimators	200

4.3. Analysis of prediction results

In this paper, MIPCA-XGBoost algorithm is applied to customer churn prediction in the telecom industry, and random forest and logistic regression algorithms commonly used in this field are used for comparison experiments. The comparison results of precision, recall and F_Score of related algorithms are shown in Table 5.

Table 5. Comparison of prediction results of various algorithms(%)

Type of Algorithm	E_{acc}	E_{recall}	F_Score
RF	84.12	84.36	83.25
LR	81.43	83.03	80.13
XGBoost	86.37	84.58	84.73
MIPCA-XGBoost	90.56	90.23	89.64

Observing the prediction results of various algorithms in Table 5, on the whole, the accuracy of three methods of random forest, logistic regression and XGBoost is more than 80%, and the MIPCA-XGBoost algorithm has the highest score, with an accuracy of 90.56%, a recall rate of 90.23%, and a F_Score value of 89.64%. The accuracy of the random forest algorithm is 84.12%, and the accuracy of the logistic regression algorithm is 81.43%, which indicates that the MIPCA-XGBoost method is more suitable for application in this field.

Secondly, through the comparison of the XGBoost algorithm MIPCA-XGBoost algorithm, it can be seen that the MIPCA feature selection method improves the prediction accuracy of the XGBoost algorithm in this experiment, its accuracy is increased by 4.19%, the recall rate is increased by 5.65%, and the F_Score value is increased by 4.91%.

5. Conclusion

From the perspective of making full use of the original information of the data set, this paper introduces the mutual information feature selection method in the study of customer churn prediction in the telecom industry, and proposes the MIPCA-XGBoost method, which successfully improves the prediction accuracy. By selecting the telecom customer data set from Kaggle website for experiments, the results show that the prediction accuracy of MIPCA-XGBoost method is as high as 90.56%. It can be seen that the algorithm is helpful for the further research of customer churn prediction in the future, and provides a new idea for enterprises to solve the problem of customer churn, helping enterprises to be more targeted and accurate in formulating customer maintenance strategies.

The research of this paper also has some shortcomings :1) At present, the rapid development of Internet communication has become an inseparable part of people's communication and contact. Under this impact, many users' communication behaviors will be affected by instant messaging tools. In the future, we will consider the third-party communication software and Internet platform data for research. 2) The data in this paper are from a single telecom industry customer data, and in the future, multiple operators in the same time period will be considered for joint research to make the research more comprehensive.

In future research, it is necessary to classify the customers who have been predicted to have churn trend, create a classification model, and then propose an effective customer retention strategy. In addition, in the study of customer churn prediction, it is necessary to pay attention to the interaction between customers, which may be an important factor affecting customer churn, which is of great significance for the research in this field.

References

[1] HADDEN J, TIWARI A, RAJKUAR R, DYMITR R. (2007) Computer assisted customer churn management: State-of-the-art and future trends[J].Computers and Operations Research, 2007, 34(10):2902-2917.

[2] ZHANG L L ,MA Y Q. (2019) Analysis of airline customer churn and consumer segmentation based on data mining algorithm using R[J].Mathematics in Practice and Theory,2019,49(06):134-142.

[3] YAN C, ZHANG X Y. (2022) Life insurance customer churn prediction algorithm based on improved K-means ang BP-Adaboost[J]. Journal of Shandong University of Science and Technology (Natural Science),2022,41(01):54-65.

[4] WU Y C. (2020) Prediction of churn rate of e-commerce customers in context of big data[J]. Modern Electronics Technique,2020,43(11):144-147.

[5] XING W, WANG S Y,ZHANG Q H, et al. (2011) Dual channel supply chain equilibrium strategy considering channer fairness[J].Systems Engineering-Theory &Practice,2011,31(07):1249-1256.

[6] LEMMENS A, CROUX C. (2005) Bagging and boosting classification trees to predict churn[J]. Journal of Marketing Research, 2005 ,43(2):276-286.

[7] FAN X L, FENG H H,YUAN M. (2013) PCA based on mutual information for feature selection[J].Control and Decision, 2013, 28(06):915-919.

[8] TAN H W, YANG Q L,XING J C, et al. (2022) Photovoltaic power prediction based on combined XGBoost-LSTM model[J].Acta Energiae Solaris Sinica,2022,43(08):75-81.

[9] PENG S Y , ZHENG G D,HUANG S J, et al. (2020) Multiple-feature short-term photovoltaic generation forecasting based on XGBoost algorithm[J].Electrical Measurement & Instrumentation, 2020,57(24):76-83.

[10] LU S, XU W M,LIU W L, et al. (2020) Short-term forecasting of PV power generation based on clustering and later regression[J].Zhejiang Electic Power,2020,39(07):48-54.

[11] HUANG C.Prediction of power genereation capacity of photovoltatic system base on artificial neural network.[D].Wuhu:Anhui Polytechnic University,2016.

[12] TANG H, WANG D,SONG B, et al. (2021) Classification of flight delay based on nonlinear weighted XGBoost[J].Journal of System Simulation,2021,33(09):2261-2269.

[13] WANG H, ZHANG W J,LIU J, et al. (2022) Flight delay prediction model based on CART algorithm[J]. Journal of Civil Aviation University of China,2022,40(03):35-40.

[14] LU M D, WEI P,HE M S,TENG Y L. (2021) Flight Delay Prediction Using Gradient Boosting Machine Learning Classifiers[J]. Journal of Quantum Computing,2021,3(1).

[15] WANG X Y, WANG Z Y,ZHAO Z, et al. (2022) A food safety risk forecast model integrated with improved AHP and XGBoost algorithm: A case study of rice[J]. Journal of Food Science and Technology,2022,40(01):150-158.

[16] MA H D. Food safety risk warning based on decision tree and random forest model[D]. Dalian: Dongbei University of Finance and Economics,2020.

[17] WANG J Y, DIAN Y F,ZHANG R F, et al. (2019) Prediction of meat product quality risk based on extreme learning machine[J]. Computer Simulation,2019,36(10):413-418.

[18] GENG Z Q, DUAN X Y,LI J T,CHU C,HAN Y M. (2022) Risk prediction model for food safety based on improved random forest integrating virtual sample[J]. Engineering Applications of Artificial Intelligence,2022,116.

[19] DING B X, ZHANG H,WANG G. (2019) Research of network intrusion detection method based on MI and SVM[J]. Journal of West Anhui University,2019,35(05):45-49+63.

[20] WANG C R, HANG D M. (2017) A study on internet customer churn prediction based on social network analysis and XGBoost [J]. Cyber Security And Data Governance, 2017, 36(23):58-61.