Journal of Intelligent & Fuzzy Systems xx (20xx) x–xx
DOI:10.3233/JIFS-190677
IOS Press

A hybrid classification model for churn prediction based on customer clustering

Qi Tang[a], Guoen Xia[a,b,*] and Xianquan Zhang[a]
[a]*College of Computer Science and Information Engineering, Guangxi Normal University, Guiling, Guangxi, China*
[b]*School of Business Administration, Guangxi University of Finance and Economics, Nanning, Guangxi, China*

Abstract. Customer churn prediction is an active research topic for the data mining community and business managers in this rapidly growing society. The ability to detect churn customers precisely is something that every company would wish to achieve. From different experiments on customer churn, it can be seen that customers always could be divided into different types and the customers in the same segment generally have similar personas, behavioral preferences, and focus points. Therefore, a hybrid classification model named ClusGBDT for customer churn prediction is proposed. This model has three steps: a feature transformation stage, a customer clustering stage, and a prediction stage. At first, the multi-layer perceptron is used to training a prediction model and replace the original attributes with low-dimensional vectors. Then, customer segments are divided using K-means. Lastly, the unique prediction model based on GBDT is constructed for every customer segment. Several measures are used to evaluate the prediction performance. From the experiments, it is observed that our design could improve original classification algorithms include random forest and logistic regression. Additionally, the proposed framework helps us to comprehend customer data.

Keywords: Customer churn, data mining, hybrid classification, customer clustering

1. Introduction

Customer churn prediction, one of the most frequently tackled tasks in customer relationship management, has gained increasing attention in recent years. It helps executives find target customers, retain customers and explore customer value [1]. Concretely, customer churn prediction is that constructing a prediction model to estimate the future churn probability for every customer using data mining technologies based on historical customer information. Generally speaking, long-term customers have stable spending power compared to new customers. And according to the survey, attracting new customers in mature markets costs several times more than to prevent regular customers from stopping services [2]. Consequently, how to formulate customer retention strategies is crucial for enterprises to enhance profitability and competitiveness.

However, constructing a credible prediction model is challenging, because the historical customer information is hidden, noisy and complicated. In previous researches, the customer churn prediction mainly has two directions. On the one hand, some researchers concentrate on improving predictive performance by constructing complex algorithms [3]. For example, support vector machine (SVM) [4–7], [28, 30] neural networks (NN) [8–10] and rough set approach [11] greatly improve the predictive performance but have difficulties to reveal the relation between churn and variables generally. On the other hand, the remaining investigators want to understand what drives customers churn from the model [12]0 so as to

*Corresponding author. Guoen Xia, School of Business Administration, Guangxi University of Finance and Economics, Nanning, Guangxi, China. E-mail: gandlf007711@163.com.

help executives make corresponding measures. For instance, decision tree (DT) [13–15] logistic regression (LR) [2, 16, 17] and random forest (RF) [16–18] have been applied to customer churn prediction because of their great robustness, comprehensibility, and great predictive performance. Until now, customer churn prediction has been widely used in various domains, including the banking sector [19–21], online gaming [22, 23], telecommunication industry [3, 24], insurance industry [25], and financial service [26].

In this paper, a hybrid predictive algorithm named ClusGBDT is proposed. This model is origin0ating from that the customers in the same segment generally have similar personas, behavioral preferences, and focus points. In our approach, customers are assigned into several segments based on customer behaviors before building classification models, which improves the performance of churn prediction and help us to comprehend the churn drives. To be specific, multi-layer perceptron (MLP) [27] is used firstly to reduce the dimension of variables for the reduction of computational cost and elimination of variable outliers. Then k-means is applied to divide customer groups and data analysis are conducted for them respectively. At last, the unique classifier based on gradient boosting decision tree (GBDT) is constructed for the corresponding customer segment.

The purposes of this study are summarized as follows:

(1) ClusGBDT is proposed as a new hybrid classification algorithm that enhances the predictive performance and robustness of GBDT based on experimental results.
(2) It helps managers to comprehend the characteristics of customers in different segments so as to formulate corresponding strategies.
(3) A general churn prediction framework for distinct industries is developed.

The rest of the paper is organized as follows: Section 2 presents the related work. In Section 3, a review of preliminaries is provided. Section 4 introduces the churn prediction model. Section 5 presents the experimental set-up. The customer segments analysis and experimental results are revealed in Section 6. The conclusions and future work are presented in Section 7.

2. Related work

2.1. Customer churn

As a result of the economic globalization and trade liberalization, a large number of companies enter the market, which results in the continuously increasing customer liquidity. Customer churn refers to the target customers who decide to abandon business services, stop purchasing products, or switch to a competitor in the market. The previous study has revealed the following three types of customer churners [11]:

(1) **Active churner:** these customers have two clear purposes, quitting the contract with the original enterprise and shifting to a competitor.
(2) **Passive churner:** these customers will not actively interrupt the business relationship with the original enterprise only if the enterprise terminates the contract.
(3) **Potential churner:** these customers will terminate the contract with the enterprise without any prior knowledge.

The first two types of churn customers can be predicted easily by manual methods. However, the third type of churn customers is difficult to predict since their historical information is extremely complicated. And the aim of customer churn prediction model is to predict the third type churner.

2.2. Review of customer churn prediction models

According to previous researches, the customer churn prediction has the following two directions.

On the one hand, some researchers concentrate on improving predictive performance by constructing complex algorithms. He et al. [28] proposed a prediction model based on SVM and random sampling. At first, random sampling is used to solve the problem of class imbalance by changing the data distribution of samples. Then, the SVM is applied to construct the prediction model. In terms of handling class imbalance, Chen et al. [29] presented a classification algorithm based on the CSCUM chart. This algorithm only needs to collect the inter-arrival time (IAT), so that the churn possibility can be estimated for the purpose of individual monitoring. Gordini et al. [30] applied SVM based on the AUC parameter-selection technique to customer churn prediction. This work showed that the process of parameter optimization plays a significant role

in prediction performance and the combination of data-driven algorithm and retention strategy is better than the common heuristic management method. Amin et al. [11] designed an intelligent rule-based approach based on four rule-generation mechanisms to extract decision rules related to churn customers, namely, Exhaustive Algorithm (EA), Genetic Algorithm (GA), Covering Algorithm (CA), and LEM2 Algorithm (LA). Stripling et al. [2] incorporated the concept of profit maximization within the customer churn prediction for the first time by using genetic algorithms to optimize the expected maximum profit measure (EMPC). Wang et al. [31] studied how the GBDT predicts the future churn possibility based on customer activities in search advertisings. This method extracts two types of features for the GBDT: dynamic features and static features at first. Then the GBDT prediction model is constructed based on these two features. Amin et al. [24] proposed a prediction method based on the distance factor which is aimed at estimating the classification certainty of different regions in the dataset.

On the other hand, some investigators want to understand what drives customer churn. Xie et al. [18] used weighted random forest (WRF) to predict churn customers. This method not only handles class balance better but also retains good interpretability. De Bock et al. [32] presented a prediction model based on rotation forest and Adaboost in 2011. The rotation forest is applied to extract customer features while the Adaboost method is used to improve predictive performance. The experimental results revealed that the predictive performance is highly improved. However, the interpretability of the model and the understandability of churn factors are deficient. Hence, De Bock et al. [33] conducted another study, combining generalized additive models (GAM) with an ensemble classification algorithm. Experimental results showed that this method not only improves the predictive performance but also has great interpretability. Verbeke et al. [34] examined the use of social network information for customer churn prediction. This method uses social network effects to handle large scale networks, a time-dependent class label, and an imbalanced class distribution. In addition, this research introduced a new method incorporating non-Markov network effects within relational classifiers and a novel parallel modeling method that combines relational and non-relational classifiers. However, the utilization rate of useful information on social networks is low in this work for three reasons. Firstly, the network characterization is tedious due to the complexity of networks and the lack of corresponding methods. Secondly, the computational cost of deriving structural features in large scale networks is high. Thirdly, most dynamic features in networks are processed as static features. Therefore, Mitrovic et al. [35] proposed a panoptic representation learning approach that integrates interactive and structural information. This approach can account for different temporal granularities by slicing the information in different periods. Yang et al. [36]0 developed a framework based on interpretable user clustering and churn prediction. This framework firstly divides users into interpretable segments, based on their daily activities and ego-network structures. Then a deep learning pipeline based on long short-term memory (LSTM) and attention mechanism is designed. Extensive data analysis and experimental results revealed that this framework helps researchers comprehend user behaviors and outperforms other prediction approaches. Caigny et al. [37]0 designed a new hybrid classification algorithm for customer churn prediction based on LR and CART which contains two stages: a segmentation stage and a prediction stage. In the segmentation stage CART is applied to identify customer segments and in the prediction stage unique models based on LR are created for all leaves of this tree.

3. Preliminaries

3.1. Notation

For computational reasons, $\boldsymbol{D}= \{(x_i, y_i) | i = 1, \ldots, m\}$ here $x_i \ \{x_i^j, j = 1, \ldots, n\}$ is assigned to denote the given dataset with m samples and n features. The binary response class is $y_i = \{1, 0\}$ signifies 'churn' if y_i takes the value 1 and 'normal' if it takes the value 0.

3.2. Review of multi-layer perceptron

The multi-layer perceptron is a traditional neural network model, as shown in Fig. 1, in which each neuron of a given layer is fully connected to all neurons of the adjacent layers, without any cycle. The training data are imported by the input layer and then processed through one or several hidden layers to compute the low-dimensional representation of customer data. At last, the output layer exports the prediction. Every neuron in the hidden layer is

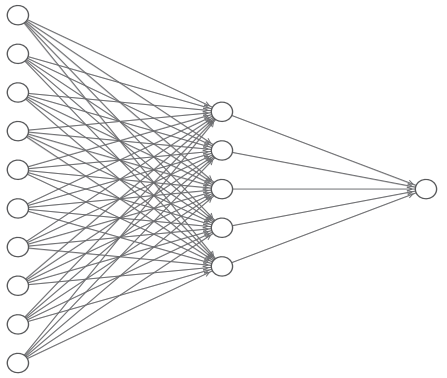


Fig. 1. Multi-layer perceptron.

consisting of the linear regression and nonlinearity as:

$$z = \sigma\left(w^T x + b\right) \tag{1}$$

where σ means the activation function. Via using backward propagation to reduce the loss function and update the parameters iteratively, the MLP could fit any classification function.

3.3. Review of K-means

K-means is a popular clustering algorithm because of its excellent comprehensibility and convergence. The goal of k-means is to find a set of centroids $C_k, k = 1, \ldots, K$ such that the total square loss is minimized:

$$\min \sum_{k=1}^{K} \sum_{i=1}^{m} ||C_k - x_i||^2 \tag{2}$$

As described in Algorithm 1, K-means alternates the optimization of centroids C_k and the assignment of each distance to the nearest centroid.

Algorithm 1 K-means

1: **Inputs:** $\{x_i\}_{i=1}^m$, K
2: Denote the label of each sample by l_i.
3: Initialize K random centroids $\{C_k\}_{k=1}^K$.
4: **repeat**
5: **for** $i = 1, \ldots, m$ **do**
6: $l_i \sim= argmin_i ||x_i - C_k||^2$;
7: **for** $k = 1, \ldots, K$ **do**
8: $C_k \sim= \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$;
9: **until** the number of changes of l_i meets some stopping criterion.
10: **return** $\{l_i\}_{i=1}^m$ and $\{C_k\}_{k=1}^K$.

4. Proposed methodology

In this section, a detailed description of the proposed model named ClusGBDT is presented step by step. The model framework is shown in Fig. 2. Our proposed model has two main differences compared to previous methods. Firstly, in contrast to most predictive methods, we construct unique models for different customer segments in parallel. Secondly, the choice of classification algorithms is not fixed in step 3. The detailed steps are as follows.

Step 1: Feature Transformation

The original customer features are This process helps us to largely reduce the computational cost and eliminate the influence of different variable distributions and useless variables. MLP has been successfully applied to customer churn prediction because of its great prediction performance. In this paper, our purpose for training an MLP model is to obtain the low dimensional representation of customer variables.

A single-layer neural network is used to construct a prediction model and the activation functions in the hidden layer are tested in the experiments including the Sigmoid function, the Tanh function, the Relu function, and the Elu function.

$$z_i = \sigma\left(\boldsymbol{W}^T x_i + b\right). \tag{3}$$

where $\boldsymbol{W}$ and b is the weight matrix and the bias vector in the hidden layer. In the output layer, the sigmoid function is used to keep the output value between 0 and 1.

$$o_i = \frac{1}{1 + \exp\left(-z_i\right)}. \tag{4}$$

If the output value o_i is more than or equal to 0.5, the model would classify this customer as a churn one. Otherwise, it means the customer is a normal one. The adaptive moment estimation (ADAM) [38] is used to optimize cross-entropy loss function:

$$j(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[y_i * \log\left(o_i\right) + \left(1 - y_i\right) * \log\left(1 - o_i\right)\right]. \tag{5}$$

where θ represents the parameters of neural networks. When the loss function converges to stopping criterion, the original feature vector of each customer is converted into a low-dimensional representation

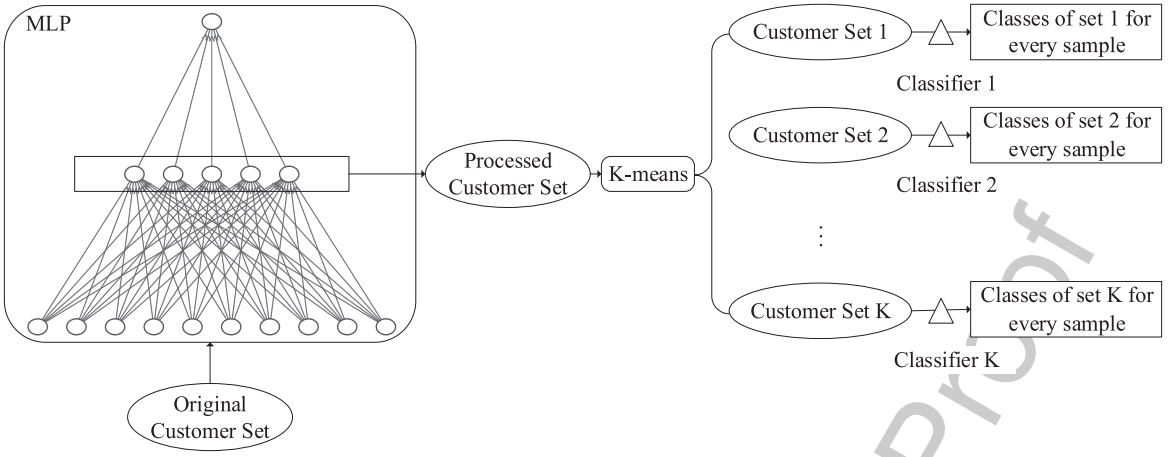


Fig. 2. Presentation of the proposed mode3.3 Review of K-means.

vector in the hidden layer of MLP.

$$t_i = z_i = \sigma\left(\boldsymbol{W}^T x_i + b\right). \tag{6}$$

where z_i is the low-dimensional representation vector in the hidden layer.

Step 2: Customer clustering

Try to use typical clustering algorithms like k-means to divide customer segments directly is challenging for three reasons:

(1) Some customer features are useless for churn prediction.
(2) Every feature in the dataset has its unique distribution.
(3) The computational cost is high when dealing with large scale datasets.

Hence, the MLP is applied to transform original features into low-dimensional presentation vectors in Step 1. This process can help us to largely reduce the computational cost and eliminate the influence of different variable distributions and useless variables.

However, how to determine the number of customer segments is challenging too. The silhouette analysis [39]0 is a popular evaluation measure of clustering performance. In this measure, $a\,(i)$ and $b\,(i)$ are used to denote the mean intra-cluster distance and the mean nearest-cluster distance respectively. The silhouette coefficient for the i-th customer is defined as:

$$s\,(i) = \frac{b\,(i) - a\,(i)}{\max\left\{\mathrm{a}\,(i)\,, b\,(i)\right\}}. \tag{7}$$

The best value is 1 and the worst value is -1. Positive values generally indicate that most samples are assigned to the right cluster. And negative values are converse. Therefore, the silhouette analysis is used to determine the number of customer segments.

The procedure of customer clustering is briefly described in Algorithm 2.

Algorithm 2 Customer Clustering

Inputs: $\{t_i\}_{i=1}^m$, K_{max}
  for $K = 1, \ldots, K_{max}$ **do**
    Denote the label of each sample by l_i.
    Initialize K random centroids $\{C_k\}_{k=1}^K$.
    repeat
      for $i = 1, \ldots, m$ **do**
        $l_i \sim= argmin_i \, ||t_i - C_k||^2$;
      for $k = 1, \ldots, K$ **do**
        $C_k \sim= \frac{1}{|C_k|} \sum_{t_i \in C_k} t_i$;
    until the number of changes of l_i meets some stopping criterion.
  Calculate the mean silhouette coefficient S_K by equation (7).
  Calculate the number of customer segments by $K = argmax\left\{S_1, S_2, \ldots, S_{K_{max}}\right\}$.
  Update the corresponding labels and centroids by K-means.
  return K, $\{l_i\}_{i=1}^m$, and $\{C_k\}_{k=1}^K$.

Step 3: Constructing a parallel prediction model

As shown in Fig. 3, we propose a parallel prediction model based on GBDT [31] 0 to make experiments. The GBDT is a high-performance method in data mining tasks, which is consisting of numerous weak prediction models.

The approximate function of GBDT is a sum of trees T_p:

$$f\,(x) = \sum_{p=1}^{P} T_p\,(x)\,. \tag{8}$$

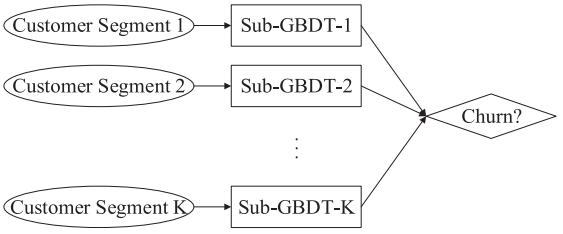


Fig. 3. Parallel GBDTs for customer churn prediction.

where P is the number of trees. The residual of p-th tree r_{ip} is denoted as

$$r_{ip} = -\frac{\partial L\left(y_i,\, f\left(x_i\right)\right)}{\partial f\left(x_i\right)}\bigg|_{f=f_{p-1}}. \tag{9}$$

Here f_{p-1} is the sum of the first p – 1 trees.

$$f_{p-1}\left(x\right) = \sum_{p=1}^{P-1} T_p\left(x\right). \tag{10}$$

Hence, the optimization objective of $T_p\left(x\right)$ is to correct its $T_{p-1}\left(x\right)$ predecessors.

$$T_p = argmin \sum_{i=1}^{m} L\left(y_i,\, f_{p-1}\left(x_i\right) + T\left(x_i\right)\right). \tag{11}$$

Note that the choice of classification algorithms is not fixed. In this paper, we use GBDT to construct the prediction model. However, other classification algorithms are appropriate for the framework.

5. Experimental setup

5.1. Experimental design

Experiments are conducted on 4 publicly available data sets from different industries. Table 1 provides an overview of these data sets and Table 2 gives some details about the features in four datasets.

In order to keep the reliability of experiments, the customer data are randomly split into training data and test data with the ratio 8:2 for 5 times and take the average performance for evaluation. All experiments

Table 1
Summary of datasets

Dataset	Industry	#samples	#numerical attributes	#discrete attributes	Churn rate (%)	Source
Ds1	Telecom	100000	77	21	49.56	Kaggle
Ds2	Music	992931	13	3	6.39	Kaggle
Ds3	Telecom	51047	34	22	28.82	Kaggle
Ds4	Financial services	10000	8	2	20.37	Kaggle

Table 2
The description of some representative features

Data set	Features	Description
Ds1	rev_Mean	Mean monthly revenue (charge amount)
	mou_Mean	Mean number of monthly minutes of use
	avg6rev	Average monthly revenue over the previous six months
	hnd_price	Current handset price
	adjqty	Billing adjusted total number of calls over the life of the customer
Ds2	payment_plan_days	Length of membership plan in days
	is_cancel	Whether or not the user canceled the membership in this transaction.
	payment_method_id	Payment method
	num_25	The number of songs played less than 25% of the song length
	num_50	The number of songs played between 25% to 50% of the song length
Ds3	Monthly Revenue	Mean monthly revenue (charge amount)
	MonthlyMinutes	Mean number of monthly minutes of use
	Total Recurring Charge	The monthly recurring charge
	Received Calls	The number of received calls
	CreditRating	A quantified assessment of the creditworthiness of a borrower
Ds4	Credit Score	The Score of credit rating
	Gender	Gender
	Num Of Products	How many accounts, bank account affiliated products the person has
	Has Cr Card	Does the customer have a credit card through the bank
	Is Active Member	Subjective, but for the concept

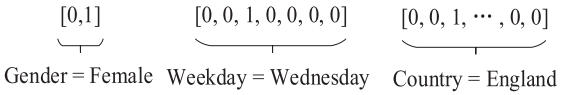


Fig. 4. One-hot encoding.

are implemented on a single machine using Sklearn and Tensorflow with a 16-core 2.1 GHz CPU (E5-2620) and a 6 Gb GPU (Quadro M5000).

5.2. Data preprocessing

In customer churn prediction, data preprocessing is a crucial step. First, the missing data in continuous variables are replaced by the mean values and then processed via z-score normalization:

$$x^j = \frac{x^j - \mu^j}{\sigma^j}. \tag{12}$$

Then, categorical variables are transformed into binary variables by one-hot encoding. This process creates v binary variables, where v means the number of distinct values including the missing value. As shown in Fig. 4, for example, one input sample [Gender = Female, Weekday = Wednesday, Country = England] is transformed into a high-dimensional sparse vector. Last, in view of some datasets are heavily imbalanced, the weighted class is applied to remedy this problem.

In addition, feature selection is a necessary component in data preprocessing. Because some variables that have lower variance are useless for classifiers and the one-hot encoding generate numerous sparse data. Hence, the features with low variance are removed to improve the predictive performance of classifiers and reduce the computational cost.

5.3. Evaluation measures

Different from the normal binary classification problem, most customer data in the customer churn prediction are extremely imbalanced. Hence, several state-of-the-art evaluation measures are used to assess the experimental results (i.e., accuracy, precision, recall, f1). In this paper, True Positive (TP) is assigned as the number of samples classified as churn customers correctly, True Negative (TN) is assigned as the number of samples classified as normal customers correctly, False Positive (FP) is assigned as the number of samples classified as churn customers falsely, and False Negative (FN) is assigned as the number of samples classified as normal customers falsely. The measures used for the evaluation of classifiers are as follows.

$$Accuracy = \frac{TP + TN}{m}. \tag{13}$$

$$Precision = \frac{TP}{TP + FP}. \tag{14}$$

$$Recall = \frac{TP}{TP + FN}. \tag{15}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}. \tag{16}$$

Additionally, we use the Receiver Operator Characteristic (ROC) curve which is a widely used metric in evaluating the performance of classifiers to present the results of experiments intuitively. The ROC curve is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. Concretely, it plots the True Positive Rate (TPR) on the x-axis and the False Positive Rate (FPR) on the y-axis.

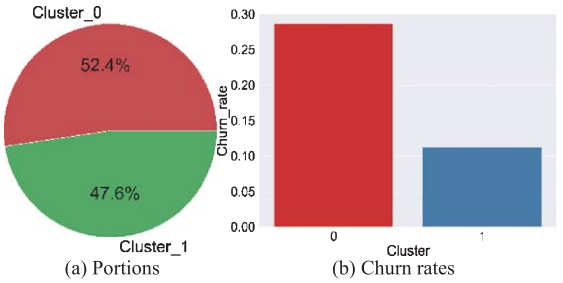
6. Results and discussion

6.1. Customer segment analysis

Dataset 4 is illustrated to analyze the customer segments for the convenience of visualization. Fig. 5 (a) presents the portions of two customer types. Additionally, the churn rates of each customer type are calculated and shown in Fig. 5 (b). The visualization shows that Cluster_0 customers are more likely to churn and the churn rate of Cluster_1 is much lower, while their counts are nearly equal. Figure 6 shows the visualization of customer segments via principal components analysis (PCA). The results intuitively show that the location between non-churn customers and churn customers is distinguishable in the same customer type. Insights like these are valuable for data analysis, customer modeling and so on.

6.2. Performance of classifiers

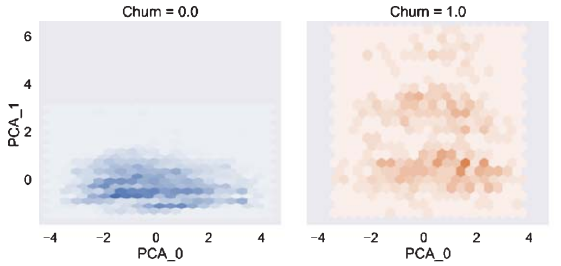
In order to demonstrate the predictive performance of our proposed model, some state-of-the-art models are applied to make comparative experiments LR [16]0, EEMLP [40], RF [18], P-MLP [10], and GBDT [31]. The brief descriptions are as follows.



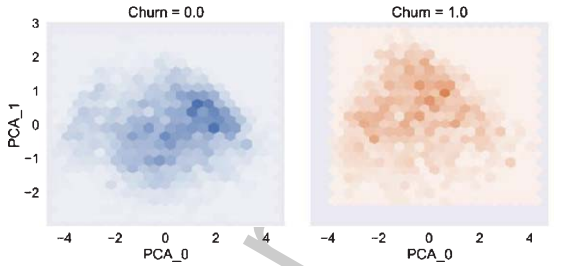
(a) Portions (b) Churn rates

Fig. 5. Portions and churn rates of two customer types.

(1) LR [16]: LR with regularization is a popular classification method for customer churn prediction due to its great timeliness and interpretability.
(2) EEMLP [40]: EEMLP is a neural network based on entity embedding. The entity embedding can efficiently force the network to learn the intrinsic properties of each feature.
(3) RF [18]: RF is a bagging ensemble model that can reduce the variance of the single decision tree.
(4) P-MLP [10]: P-MLP is a particle classification optimized-based neural network which has a lower learning error and s faster convergence speed than traditional networks.
(5) GBDT [31]: GBDT is boosting ensemble model. IN contrast to RF, GBDT also aims to minimize the bias and not only the variance.



(a) The PCA of Cluster_0



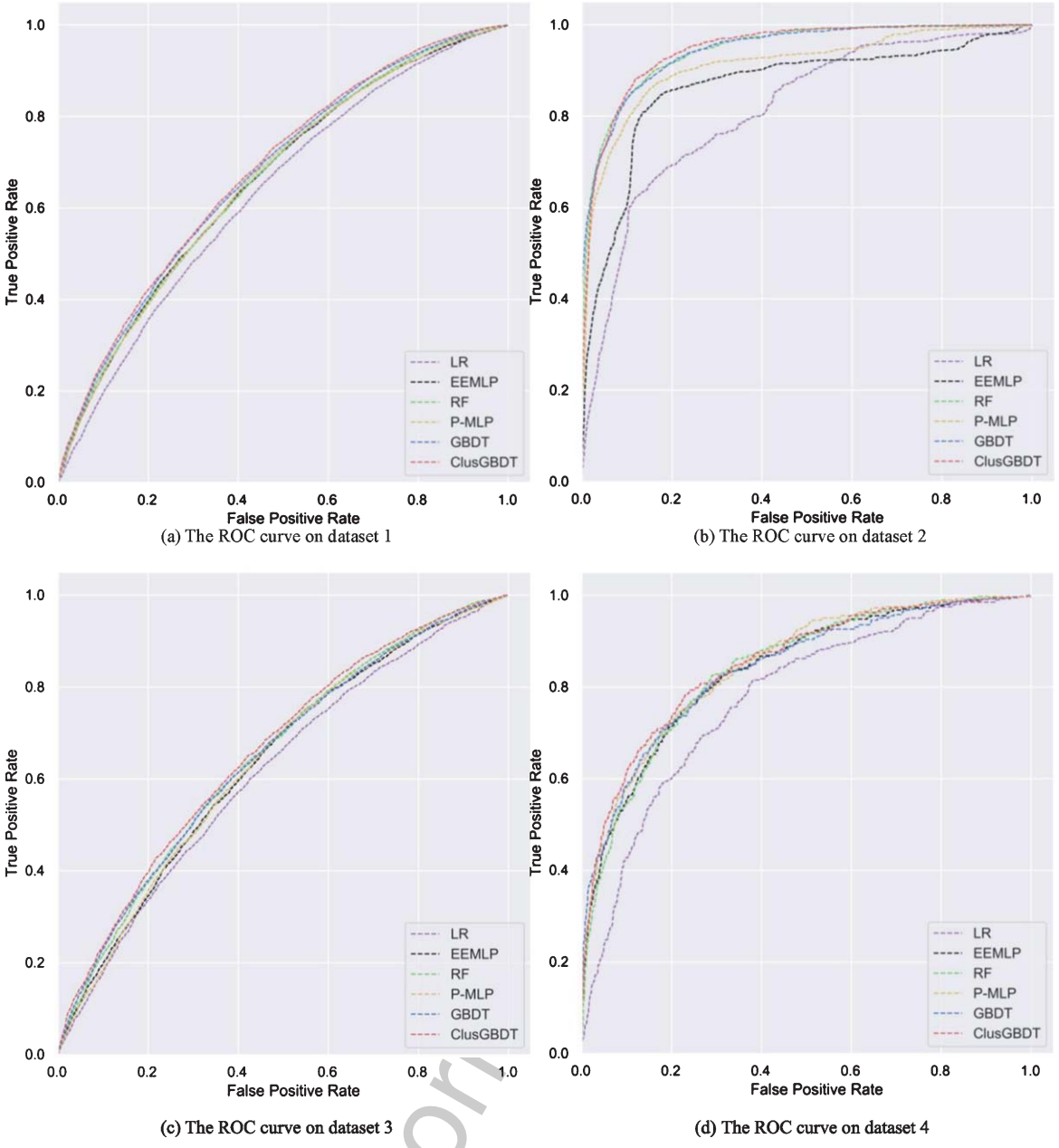
(b) The PCA of Cluster_1

Fig. 6. The PCA of customer segments.

The average cross-validation predictive performance of different models over four datasets are listed in Table 3. The best performance classifier in each dataset is in bold. Additionally, the ROC curves are shown in Fig. 7. These average results are the basis of a statistical analysis of model performance. From

Table 3
The prediction performance of several models

Data set	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Ds1	LR	59.47	58.83	60.68	59.81
	EEMLP	61.71	61.10	62.58	61.83
	RF	61.52	60.32	**65.34**	62.73
	P-MLP	61.67	61.13	62.22	61.67
	GBDT	63.08	62.56	62.78	62.67
	ClusGBDT	**63.21**	**62.74**	64.15	**63.44**
Ds2	LR	87.29	27.83	62.07	38.43
	EEMLP	87.26	31.67	82.44	45.76
	RF	87.55	32.58	**88.52**	47.63
	P-MLP	87.89	32.77	81.60	46.76
	GBDT	90.02	37.89	83.05	52.04
	ClusGBDT	**90.03**	**38.72**	83.67	**52.94**
Ds3	LR	58.33	35.74	55.93	43.61
	EEMLP	61.13	38.45	58.12	46.29
	RF	59.26	37.57	62.50	46.93
	P-MLP	**62.28**	38.93	54.35	45.37
	GBDT	60.38	38.33	61.59	47.26
	ClusGBDT	60.74	**39.08**	**64.70**	**48.71**
Ds4	LR	70.70	37.86	68.45	48.74
	EEMLP	82.64	56.44	65.21	60.48
	RF	77.03	45.97	**73.07**	56.42
	P-MLP	82.79	57.46	59.21	57.98
	GBDT	83.44	58.98	61.47	60.16
	ClusGBDT	**84.20**	**63.64**	61.95	**62.49**



(a) The ROC curve on dataset 1

(b) The ROC curve on dataset 2

(c) The ROC curve on dataset 3

(d) The ROC curve on dataset 4

Fig. 7. The ROC curves on four public datasets.

the experiments, our proposed model performed better than other algorithms in most evaluation measures. And further researches on predictive performance are discussed in section 6.3.

6.3. Discussion of results

First, we focus on dataset 1 (a balanced dataset), the ClusGBDT outperformed the other five methods that proves our proposed method has a great binary classification ability. As to the other three datasets, the metrics: Precision, Recall, and F1-score are much more important because these datasets are imbalanced. When dealing with imbalanced problems, the ability to learn the features of churn customers is significant for a classification model. Similarly, the ClusGBDT performed much better. Different from traditional evaluation measures, the ROC curve can express dynamic predictive performance as its discrimination threshold is varied. We can intuitively

Table 4
The impact of customer clustering on LR and RF

Data set	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Ds1	LR	59.47	58.83	**60.68**	59.81
	ClusLR	**60.05**	**59.61**	60.08	**59.82**
	RF	61.52	**60.32**	65.34	62.73
	ClusRF	**61.65**	60.29	**65.46**	**62.77**
Ds2	LR	**87.29**	27.83	62.07	38.43
	ClusLR	86.47	**28.74**	**71.66**	**40.96**
	RF	87.55	32.58	88.52	47.63
	ClusRF	**87.61**	**35.25**	**88.78**	**50.16**
Ds3	LR	58.33	35.74	55.93	43.61
	ClusLR	**59.24**	**36.45**	**56.11**	**44.17**
	RF	59.26	37.57	**62.50**	46.93
	ClusRF	**60.62**	**38.32**	61.01	**47.07**
Ds4	LR	70.70	37.86	68.45	48.74
	ClusLR	**73.73**	**40.80**	**71.16**	**51.30**
	RF	77.03	45.97	**73.07**	56.42
	ClusRF	**79.63**	**48.97**	70.31	**57.47**

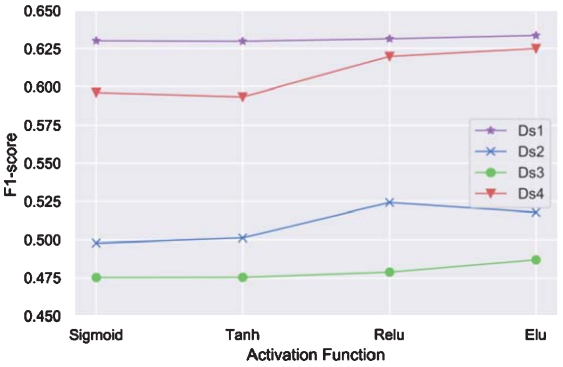


Fig. 8. The influence of activation function on F1-score.

observe that our proposed method has a superior performance. To sum up, the ClusGBDT outperforms other predictive methods and has a great robustness no matter whether the churn rate is low or high and the amount is large or small.

6.4. *The influence of activation function*

We study the impact of activation functions on ClusGBDT. In this section, the experiments are conducted via holding the other settings. Note that we exploit the identity as activation function on neurons of MLP, as shown in Equation (3). A common practice in neural networks is to test non-linear activation functions on hidden layers. The choice of activation functions is related to the accuracy of customer segments. We thus compare the F1-score of different activation functions on ClusGBDT. The experimental results are shown in Fig. 8.

6.5. *The influence of customer clustering for LR and RF*

In order to study the impact of customer clustering for other classifiers in depth, we also deploy customer clustering on LR and RF that are frequently used in the domain of customer churn prediction. As shown in Table 4, the improvements for LR and RF are obvious on most evaluation measures. In general, customers in the same segment always have similar personas, behavioral preferences, and focus points. It helps to detect exactly the churn drives for each and every segment. Afterwards, the managers can take appropriate strategies for every segment and tackle the churn drives of the segment. And the classification techniques can learn the customer features better to improve predictive performance.

7. Conclusion and future work

In this study, the application of MLP is explored to customer clustering for churn prediction. We aim to ensure the accuracy and reliability of customer segments. MLP can enhance the interaction between customer features and eliminate the influence of useless features without any feature selection, extraction, and generation. To evaluate the performance of the proposed model, five algorithms are compared and several datasets are tested. In our benchmarking research, the proposed model is the overall most great compared to other classification techniques. To sum up, our model offers three main contributions to the existing literature such as: (1) improved the

prediction performance and robustness of traditional algorithms, (2) help managers to comprehend the causes of customer churn so as to formulate corresponding strategies, and (3) developed a general churn prediction framework for distinct industries.

As a topic of further research, the meaning of high-order vectors is needed to study so that the managers could formulate corresponding strategies for several customer segments. Another future direction is to improve the efficiency of the model. Because the training time spending on the proposed model is too much. The model training time is also a necessary metric to take into consideration especially in cases where real-time predictions are needed. Lastly, the predictive performance of our proposed framework is restricted to the performance of MLP. Hence, it is significant to find a way to ensure the accuracy of customer segments.

Acknowledgments

Our work is partially supported by National Natural Science Foundation of China (No. 71862003), the Foundation of Guangxi Key Laboratory Cultivation Base of Cross-border E-commerce Intelligent Information Processing, Guangxi University of Finance and Economics, and the Foundation of Science of Business Administration, Guangxi University of Finance and Economics.

References

[1] A. berson, S. Smith and K. Thearling, Building data mining applications for CRM, McGraw-Hill, New York, (1999).

[2] E. Stripling, S. Broucke, K. Antonio, B. Baesens and M. Snoeck, Profit maximizing logistic model for customer churn prediction using genetic algorithms, *Swarm and Evolutionary Computation* **40** (2018), 116–130.

[3] W. Verbeke, K. Dejaeger, D. Martens, J. Hur and B. Baesens, New insights into churn prediction in the telecommunication sector: A profit driven data mining approach, *European Journal of Operational Research* **218**(1) (2012), 211–229.

[4] H. Kaizhu, D. Zheng, J. Sun, Y. Hotta, K. Fujimoto and S. Naoi, Sparse learning for support vector classification, *Pattern Recognition Letters* **31**(13) (2010), 1944–1951.

[5] I. Brandusoiu and G. Toderean, Churn prediction in the telecommunications sector using support vector machines, *Annals of The Oradea University* **1** (2013), 19–22.

[6] J. Runge, P. Gao, F. Garcin and B. Faltings, Churn prediction for high-value players in casual social games, in: *IEEE Conference on Computational Intelligence and Games*, Washington DC, (2014).

[7] M. Farquad, V. Ravi and S.B. Raju, Churn prediction using comprehensible support vector machine: An analytical CRM application, *Applied Soft Computing* **19** (2014), 31–40.

[8] P.C. Pendharkar, Genetic algorithms based neural network approaches for predicting churn in cellular wireless network services, *Expert Systems with Applications* **36**(3) (2009), 6714–6720.

[9] H. Faris, Neighborhood cleaning rules and particle swarm optimization for predicting customer churn behavior in telecom industry, *International Journal of Advanced Science and Technology* **68** (2014), 11–22.

[10] R. Yu, X. An, B. Jin, J. Shi, O.A. Move and Y. Liu, Particle classification optimization-based BP network for telecommunication customer churn prediction, *Neural Computing and Applications* **29** (2018), 707–720.

[11] A. Amin, S. Anwar, A. Adnan, M. Nawaz, K. Alawfi, A. Hussain and K. Huang, Customer churn prediction in the telecommunication sector using a rough set approach, *Neurocomputing* **237** (2017), 242–254.

[12] H. Hansen, B.M. Samuelson and J.E. Sallis, The moderating effects of need for cognition on drivers of customer loyalty, *European Journal of Marketing* **47**(8) (2013), 1157–1176.

[13] C. Kirui, L. Hong, W. Cheruiyot and H. Kirui, Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining, *International Journal of Computer Science Issues* **10**(2) (2013), 165–172.

[14] E. Shaaban, Y. Helmy, A. Khedr and M. Nasr, A proposed churn prediction model, *International Journal of Engineering Research and Applications* **2**(4) (2012), 693–697.

[15] S.A. Qureshi, A.S. Rehman, A.M. Qamar, A. Kamal and A. Rehman, Telecommunication subscribers' churn prediction model using machine learning, in: *Proceedings of the Eighth International Conference on Digital Information Management*, (2013), pp. 131–136.

[16] K. Coussement, S. Lessman and G. Verstraeten, A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication, *Decision Support Systems* **95** (2017), 27–36.

[17] S. Neslin, S. Gupta, W. Kamakura, J. Lu and C. Mason, Detection defection: Measuring and understanding the predictive accuracy of customer churn models, *Journal of Marketing Research* **43**(2) (2006), 204–211.

[18] Y. Xie, X. Li, E.W.T. Ngai and W. Ying, Customer churn prediction using improved balanced random forests, *Expert Systems with Applications* **36**(3) (2008), 5445–5449.

[19] K. Chitra and B. Subashini, Customer retention in banking sector using predictive data mining technique, in: *The 5th International Conference on Information Technology*, (2011), pp. 1–4.

[20] U.P. Devi and S. Madhavi, Prediction of churn behavior of Bank customers using data mining tools, *Indian Journal of Marketing* **5**(1) (2012), 96–101.

[21] A.O. Oyeniyi and A.B. Adeyemo, Customer churn analysis in banking sector using data mining techniques, *African Journal of Computing and ICT* **8**(3) (2016), 165–174.

[22] J. Kawale, A. Pal and J. Srivastava, Churn prediction in MMORPGs: A social influence based approach, in: *Proceedings 2009 International Conference Computational Science and Engineering* **4** (2009), 423–428.

[23] M. Suznjevic, I. Stupar and M. Matijasevic, MMORPG Player behavior model based on player action categories, in: *The 10th Annual Workshop on Network and Systems Support for Games*, IEEE Press, 2011.

[24] A. Amina, F. Al-Obeidatb, B. Shah, A. Adnan, J. Loo and S. Anwar, Customer churn prediction in telecommunication

industry using data certainty. *Journal of Business Research* **94** (2019), 290–301.

[25] R.A. Soeini and K.V. Rodpysh, Applying data mining to insurance Customer churn management, in: *International Proceedings of Computer Science and Information Technology* **30** (2012), 82–92.

[26] D. Van den Poel and B. Larivière, Customer attrition analysis for financial services using proportional hazard models, *European Journal of Operational Research* **157**(1) (2004), 196–217.

[27] H. Ramchoun, M.A.J. Idrissi, Y. Ghanou and M. Ettaouil, Multilayer perceptron: Architecture optimization and training, *International Journal of Interactive Multimedia and Artificial Intelligence* **4** (2016), 26-30.

[28] B. He, Y. Shi, Q. Wan and X. Zhao, Prediction of customer attrition of commercial banks based on SVM model. *Procedia Computer Science* **31** (2014), 423–430.

[29] S.H. Chen, The gamma CUSUM chart method for online customer churn prediction. *Electronic Commerce Research and Applications* **17** (2016), 99–111.

[30] N. Gordini and V. Veglio, Customer churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry, *Industrial Marketing Management* **62** (2017), 100–107.

[31] Q.F. Wang, M. Xu and A. Hussain, Large-scale ensemble model for customer churn prediction in search ads, *Cognitive Computation* **11** (2019), 262–270.

[32] K.W. De Bock and V.D. Poel, An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction, *Expert Systems with Applications* **38**(10) (2011), 12293–12301.

[33] K.W. De Bock and V.D. Poel, Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models, *Expert Systems with Applications* **39**(8) (2012), 6816–6826.

[34] W. Verbeke, D. Martens and B. Baesens, Social network analysis for customer churn prediction, *Applied Soft Computing* **14** (2014), 431–446.

[35] S. Mitrovic, B. Baesens, W. Lemahieu and J.D. Weerdt, tcc2vec: RFM-informed representation learning on call graphs for churn prediction, *Information Sciences*, (2019), Available online.

[36] C. Yang, X. Shi, J. Luo and J. Han, I know you'll be back: Interpretable new user clustering and churn prediction on a mobile social application, in: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2018), pp. 914–922.

[37] A.D. Caigny, K. Coussement and K.W. De Bock, A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees, *European Journal of Operational Research* **269** (2018), 760–772.

[38] D.P. Kingma and J.L. Ba, ADAM: A method for stochastic optimization, in: *The 3rd International Conference for Learning Representations*, (2015).

[39] R.J.G.B. Campello and E.R. Hruschka, A fuzzy extension of the silhouette width criterion for cluster analysis, *Fuzzy Sets and Systems* **157**(21) (2006), 2858–2875.

[40] C. Guo and F. Berkhahn, Entity embedding of categorical variables, *arXiv Preprint arXiv: 1604.06737*, 2016.