The Study of Customer Split-Flow Based on Data Mining

CHEN Jun-De¹ and ZHANG De-Fu²

¹Huawei Technologies Co.Ltd, Shenzhen, Guangdong, China
²School of information science and engineering, Xiamen University, China

chen2wo@126.com

Abstract. In this paper, we established the business preference model to identify who often applied the low-value businesses in service hall, and the channel adapter model was built, which was used by data mining, and we made this model determine the suitability customers for them to apply the low-value business by electronic channels. Three categories were divided that fit the specific recommendation, required patience to guide, and not suitable recommended. Finally, according to the actual business data, we validated the model and obtained the satisfactory results for customer split-flow.

1. Introduction
With the growth of the business for people who handle their affairs in physical business hall and hotline channel, that is necessary to divert low-value business users to other channel such as electronic business hall, induce them to handle the business in online business hall directly and not crowd the high value business in physical hall.

Data mining is a new technology to meet the needs of extracting information from massive databases in information society. It can help decision-makers extract the implicit, unknown, non trivial and potentially valuable information or knowledge [1]. Classification problem is the critical problem, and classification algorithm is the key technology for data mining, it has a wide application in many fields, such as pattern recognition, data analysis, image processing and so on. Such as Sujinlong, et al. [2] pretreated a video image, and used the method of fuzzy neural network to implement target identification. Based on the customer data of game, Tzu-Liang (Bill) Tseng, et al. [3] classified the customer. Michael Y. Hu et al. [4] extracted the features by the method of neural network etc. Classification problem is becoming a very active research topic in the field of data mining, and this split-flow for business hall could be regarded as the classification problem as also.

For the split-flow, the customers should be classified and suitable customer was required to be distinguished firstly. By applied the classification algorithm of data mining, we could distinguish and perform the split-flow for the users who often deal with the existing physical hall and hotline channels for the low value business. From three aspects of customer, business, channel, the three-dimensional adaptation model was built, which could reduce the operating pressure in physical channel, and increase the service efficiency in electronic channels, improve the user satisfaction.

2. Frequent Classification Algorithm

2.1. FRI Algorithm
Fuzzy Rules Induction (FRI) is a self-organizing data mining and non-parametric method based on Group Method of Data Handling (GMDH) [5]. FRI algorithm could automatically extract the fuzzy rules from data, and describe the complex systems in the form of IF-THEN by close natural language (such as good, general, poor). FRI Uses the black-box approach to analyze and dispose the relation of input and output variables in system, and automatically extract the fuzzy rules from the data by GMDH technology, this algorithm could give a forecast for the system, and is a kind of effective way to describe the behavior of the system.

2.2. Artificial Neural Network (ANN)

2.2.1. ANN Algorithm Principle. Artificial neural network (ANN) is a new intelligent handling method in the process of imitating human brain to solve problem, through a lot of simple handling units called neurons constitute, it constitutes a nonlinear dynamic system, and the basic neuron model is shown as Figure 1.

![Artificial Neuron](image)

**Figure 1. Artificial Neuron**

2.2.2. ANN Classification. The fuzzy rules of if-then form are the most basic knowledge of complex decision-support techniques. The classical theoretical methods for obtaining fuzzy rules include the rough set theory proposed by Pawlak and the fuzzy model proposed by Takagi and Sugeno[6], ChenHui [7] proposed that the rough sets and fuzzy sets has certain limitations, such as fuzzy set has unthoroughness, and rough sets has generality, etc. HanMin, et al. [8] proposed a classification method for multi-attribute fuzzy rules based on cloud-neural network, based on these analyses the rule classification method of neural network was studied in the paper. For discrete properties, the front parts of the rule classification are usually represented as Boolean functions. For example, the sample has two discrete properties \((x_1, x_2)\), and the value of \(x_1\) maybe be \(x_{11}, x_{12}, x_{13}, x_{14}\); The possible value of \(x_2\) maybe be \(x_{21}, x_{22}\). there are 2*4 species, and the format of the extracted classification rules is shown below.

\[
IF(x_1 = x_{11}) \lor (x_1 = x_{12} \land x_1 = x_{13}) \lor (x_1 = x_{13} \land x_1 = x_{14}) \lor (x_1 = x_{14}) \\
THEN t = 1 \quad ELSE \quad t = 0
\] (1)

2.3. Decision Tree

The common algorithms of decision tree are ID3 and C4.5, the C4.5 algorithm is based on the ID3 algorithm and proposed by Quinlan in 1993. As a typical classification algorithm, decision tree[9] has been used in the financial, medical and other industries widely. Similar to the above algorithm, the data pretreatment should be performed before the compute of decision tree, the continuous attribute was needed to be discretized, and the training set should be formed for the decision tree, and the basic theory of C4.5 algorithm is clear and simpler.
3. Algorithms Selection

3.1. FRI Algorithm

So, through analyze the algorithm principle and characteristics for artificial neural network, decision tree, FRI, we have a basic understanding for the advantages and disadvantages of each algorithms. It should be pointed out that, in view of the different problems or scene, there will be different method to solve it, which requires a scientific approach to analyze and study. When faced the general classification problem, we would estimate which kind of algorithms are done relatively well or suitably in advance.

Therefore, the use of test and comparison would be a feasible method to select the suitable algorithm, and this could be implemented well in the UCI data sets. Took the balance-scale dataset for example, we selected 500 samples orderly to constitute the training set, and 125 samples to constitute the test data. Used the above algorithms to calculate respectively, we acquired the average error rate was 9.4% in training set, and 6.4% in test set. Then, we chose other part respectively for testing set, and the rest of the four parts as a learning set, this dataset was run repeatedly five times. The average error rate in the test set is 13.44%. The extraction rule for L class is

\[ IF \ \text{NOT} - A, \ & \text{OR} \ \text{NOT} - B, \ & \text{THEN} \ L \]

The signification of above expression (2) is that if the weight both the left and right side is not 1, or the distance between left and right is not 1, then it belongs to L class. Similarly, the Vote dataset has 435 samples, we constituted the 348 samples as the learning set, 87 samples as the test set, we acquired that the average error rate in the test set is 3.45%.

The other data sets were processed similarly, by contrasted the test results for the above algorithms of FRI, decision tree and ANN, we could get the different accuracy of these algorithms and select the relatively good algorithm to apply. In this case, we selected the decision tree to distinguish the customer and implement the business split-flow.

4. Model Building

4.1. Channel adaptation model

4.1.1. Model Preparation. Both the domestic and foreign scholars have some study in the channel preference of customer. Such as Siwon Cho, et al.[10] studied the preference of physical or virtual channel for the customers. Chiang, et al.[11] used the artificial neural network to predict the customers’ selection behavior for different channels. Kevin, et al.[12] first used AHP to estimate the customers’ channel preference, etc.

The Channel optimization analysis should firstly analyze the urgent split-flow business to confirm the target customer group, and on this basis, the channel adaptation model would be built to determine whether the customer was worth to recommend or not, through the relevant measures were taken to appropriate customer, the goal of improving customer satisfaction was able to be achieved. Through the establishment of the customer - business - channel adaptation model, it helps the manager take pointed split-flow measures to target customer groups and guide the customers to use remote electronic channels.

As is shown in the below table 1, according to the business value, the customer business was divided into three levels of high, middle and low, and at the same time, the relevant customer channel was divided into three categories of Suitable recommendation, Need to be patient guidance and not Suitable recommendation as also. The detail was as follows.

| Customer-business | Customer-channel | measures and strategies |
|-------------------|------------------|------------------------|
| High              | Suitable recommendation | Priority diversion guidance |
High  Need to be patient guidance  Suggest manual or SMS recommendation
High  Not suitable for recommendation  Temporarily not diverting
Median  Suitable recommendation  Priority diversion guidance
Median  Need to be patient guidance  Suggest manual or SMS recommendation
Median  Not suitable for recommendation  Temporarily not diverting
Low  Suitable recommendation  Priority diversion guidance
Low  Need to be patient guidance  Suggest manual or SMS recommendation
Low  Not suitable for recommendation  Temporarily not diverting

4.1.2. Data Preprocessing. Through the comprehensive analysis for the data of customer channel, the key fields of the channel adaptation model are determined, including users age, type, Average electricity fee for nearly three cycle, Number of arrears, Transfer failure times within a year, Cash payment times within a year, contract content, whether the electronic bill was opened, the IVR times within a year, total 9 fields were extracted and these fields were represented by $X_i (i = 1, 2, 9)$ respectively. The extracted fields were calculated twice, and the data was preprocessed at the same time, including the missing value processing, abnormal value processing, data regulation and dimension reduction, etc.

4.1.3. Model Process. By analyzed the proportion of channel business, the service staff could obtain which was the most urgent business type for the channel split-flow. The business preference model was used to identify who often applied the low-value businesses in service hall, and the channel adapter model was used to channel split-flow, such as the low-value business was guided to electronic channels and so on.

As the chart shows below, the channel adapter model and business preference model were built separately. Among the target customers, the channel adaption model was calculated. The customer channel was divided into three categories, and each category was recommended to the specific channel. The customer - business preference model was calculated as also, and the prediction probability was calculated for the business preference. Then, we used the model in the actual business data and verified it. The last was the result analysis. Detail was listed as Figure 2.

![Figure 2. The Chart of Modeling Process.](chart)

In this way, the model could distinguish the customer who transacted low-value business in physical business hall and hotline channel. Here, the algorithm of decision tree was used in the model, and the typical algorithm of decision tree is ID3 and C4.5. The C4.5 algorithm was improved from ID3 and inherited the advantages of ID3 algorithm. Compared with other classification algorithms,
such as statistical methods and neural networks, the C4.5 algorithm had the following advantages: the generated classification rules are easy to understand and the accuracy is higher.

The C4.5 decision tree model was used to distinguish whether the customers were suitable to transact the business by electronic channels or not, and the calculation process was shown in Figure 3. Then, it will be guided and recommended to the electronic channels for the suitable user categories, in this way, the purpose of channel diversion was achieved.

![Figure 3. C4.5 decision tree](image)

$L$ is the final leaf node, the recommended proportion of each leaf node is the probability fit to recommend the electronic channel, below table 2 is the detailed result

| USERS | Attr | Suitable recommendation | Need patient guidance | Not suitable recommendation |
|-------|------|-------------------------|-----------------------|-----------------------------|
| USER 1 |      | 87%                     | 13%                   | 0%                          |
| USER 2 |      | 70%                     | 25%                   | 5%                          |
| USER 3 |      | 20%                     | 78%                   | 2%                          |
| USER 4 |      | 0%                      | 17%                   | 83%                         |

The classification categories include the recommended for electronic channels, need patience guidance of electronic channels, and not suitable to recommend for electronic channels, at the same time, the probability of each category was calculated as above.

As to the different channel, different customer has the different preference for the channels, and table 3 is the result of preference for different customer channel. Such as the large users, this group was favor of the experience platform, and the most of them would like to transact the business by experience platform, the probability is 47%.  

5
Table 3. Calculation Result of Customer Preference

| Matrix of Customer-Channel | BOSS  | Self | coop | SMS  | Online sales | Experience platform | Hotline |
|----------------------------|-------|------|------|------|--------------|---------------------|--------|
| Large users                | 51%   | 51%  | 41%  | 35%  | 47%          | 84%                 |
| Family users               | 29.5% | 29%  | 12.5%| 8.5% | 31.5%        | 56%                 |
| Indiv-com users            | 25.9% | 30%  | 10.8%| 9.5% | 34%          | 70%                 |

5. Conclusion
The model of customer split-flow is established to divert customers' pressure on reducing physical channel operations, improve overall service efficiency and enhance customer satisfaction. It consists of two parts, one is the channel adaptation model, and the other is the business preference model. First of all, the customers who often deal with the low value business are identified, and according to the business preference model, the customers are divided into three categories of high, medium and low as also. Then, the channel adaptation model is established to determine whether the customers are suitable to electronic channels or not, this divided categories are suitable for recommendation, need for patient guidance and not recommended. Finally, the model is validated on the actual business data, and the satisfactory split-flow results are obtained.

6. Acknowledgment
The authors wish to thank the project chance provided by Guangzhou Power Supply Bureau, and thank Mr. Zhang Liang-jun, the chairman of Guangzhou TipDM Intelligent Technology Co., Ltd., for valuable discussion and contribution to the successful delivery of the project.

References
[1] Zhang Liangjun, Chen Junde, Liu Mingjun, Chen Rong. Data mining: practical case study [M]. Beijing: China Machine Press, 2013: 10-1
[2] Su kanadaki, HERBERT H C IU, FERNANDO T, Embedded Design in Neural Network and Optical Flow Based High-speed Target Recognition & Tracing [J]. China Journal of Highway and Transport 2015, 28 (11): 113 – 121
[3] Tzu-Liang(Bill)Tseng, Chun-CheHuang, Rough set-based approach to feature selection in customer relationship management [J]. Omega 35 (2007) 365–383.
[4] Michael Y. Hu , Murali Shanker , G. Peter Zhang, Ming S. Hung Modeling consumer situational choice of long distance communication with neural networks[J]. Decision Support Systems 44 (2008) 899–908
[5] He Changzheng. Self-organizing data mining and economic forecasting [M]. Beijing: Science Press, 2005.
[6] Leng G, McGinnity T M, Prasad G. An approach for on-line extraction of fuzzy rules using a self-organising fuzzy neural network[ J]. Fuzzy Sets and Systems,2005, 150(2): 211-243.
[7] ChenHui, Li De-Yi, Shen Cheng-Zhi. A Cloud Model Applied to Controlling inverted pendulum [J]. Journal of Computer Research& Development, 1999, 36 (10): 1180-1187.
[8] HanMin,LiZheng.Multi-attribute fuzzy rules classification based on cloud-neural network[J].Control and Decision,2009, 24(6): 933-936.
[9] Lin Zhang , et al. Efficient privacy-preserving classification construction model with differential privacy technology[j]. Journal of Systems Engineering and Electronics,2017(01)
[10] Siwon Cho, Jane E. Workman Effects of need for touch, centrality of visual product aesthetics and gender on channel preference for apparel shopping[J]. Journal of Global Fashion Marketing, 2015, Vol. 6(2), pp. 120-135.

[11] Chiang W Y K, Zhang D, Zhou L. Predicting and Explaining Patronage Behavior toward Web and Traditional Stores Using Neural Networks: A Comparative Analysis with Logistic Regression[J]. Decision Support System, 2006, 42(2):18.

[12] Kevin Chiang, Zhen Li. An analytic hierarchy process approach to assessing consumers’ distribution channel preference [J]. International Journal of Retail & Distribution Management, 2010, Vol. 38(2) 2010.