Customer Value Analysis for Telecom Industry Based on Data Mining

Xiaoyong Liu and Hui Fu

***Abstract*—Customer plays an important role in survival and development of telecom enterprises. Every customer claims different development and care costs, which requires telecom enterprises to make different resource investment programs to customers of different types of value. Customer value analysis is necessary in order to guide enterprises to make optimum resource allocation effectively. In this paper, a model used to evaluate comprehensive value of customers was established according to the customer value evaluation system, and an empirical analysis based on real customer data of a telecom enterprise was implemented.**

***Index Terms*—Data mining, comprehensive evaluation system, customer value, principal component analysis (PCA).**

I. Introduction

With the technological development and social progress, the infrastructure gap in the telecom industry is narrowing gradually and the scope of business tends to be identical. Facing with the increasing fierce competition in the telecom industry, product innovation alone is far from enough to maintain the dominant position in the industry [1]-[4]. Therefore, telecom enterprises shall attach high attentions on how to retain high-value customers and develop great potential ones with limited resources to achieve the maximum profit and sustainable development. As the analysis hotspot in recent years, data mining achieves rapid development and has been widely applied in different fields. By analyzing real customer data of telecom enterprises, this paper put forward an evaluation model of comprehensive customer value based on the customer value evaluation system [5]-[8].

II. Customer Value Evaluation Model for Telecom Enterprises

A. Establishment of the Customer Value Evaluation Model

1) Establishment of analytic hierarchy model [9], [10]

As listed in Table I, customers were firstly divided according to current value and potential value. Analytical hierarchy process was conducted to current value and potential value, respectively. The goal layer is divided into Current Value A_1 and Potential Value A_2; the criterion layer is divided into Profit Contribution B_1, Cost B_2 and Satisfaction B_3; the index layer is divided into ARPU C_1, Clearing Fee C_2, Call Proportion in Busy Hours C_3, Call Proportion in Idle Hours C_4, Cost Subsidies C_5, Preferential Fee C_6, In-the-net Time C_7, Long-distance and Roaming Ratio C_8, Value-added Service Penetration Rate C_9 and Data Traffic Saturation C_{10}.

TABLE I: Classification of Indexes

Goal layer	*Criterion layer*	*Index layer*
Current Value	Profit Contribution	ARPU
	Cost	Clearing Fee
		Call Proportion in Busy Hours
		Call Proportion in Idle Hours
		Cost Subsidies
		Preferential Fee
Potential Value	Satisfaction	In-the-net Time
		Long-distance and Roaming Ratio
		Value-added Service Penetration Rate
		Data Traffic Saturation

The pair wise comparison matrix of Profit Contribution and Current Value is: $\begin{bmatrix} 1 & 2 \\ 1/2 & 1 \end{bmatrix}$.

The eigenvector corresponding to the largest eigenvalue of the matrix was calculated through the Matlab software: $[0.8944, 04472]^T$. Normalize it and then the Profit Contribution could be gained. Weight of the Profit Contribution to the Current Value is: $[0.67, 0.33]^T$

Consistency test:

$$CI = \frac{\lambda - n}{n-1} = \frac{2-2}{2-1} = 0$$

$$CR = \frac{CI}{RI} = 0 \leq 0.1$$

If CR=0< 0.1, it is viewed passing the consistency test. Therefore, weight of the Profit Contribution to the Current Value is: (0.67, 0.33).

Similarly, pairwise comparison matrices of Clearing Fee, Call Proportion in Busy Hours, Call Proportion in Idle Hours, Cost Subsidies and Preferential Fee to the Cost were constructed.

$$\begin{bmatrix} 1 & 3 & 3 & 2 & 2 \\ 1/3 & 1 & 1 & 1/3 & 1/3 \\ 1/3 & 1 & 1 & 1/3 & 1/3 \\ 1/2 & 3 & 3 & 1 & 2 \\ 1/2 & 3 & 3 & 1/2 & 1 \end{bmatrix}$$

Manuscript received August 9, 2015; revised March 2, 2016.

The authors are with the Department of Computer Science, Guangdong Polytechnic Normal University, Guangdong, 510665, China (e-mail: lxyong420@126.com, lindafh819@126.com).

The eigenvectors corresponding to the largest eigenvalues of the matrices were calculated through the Matlab software: $[0.704, 0.174, 0.174, 0.532, 0.402]^T$.

Weights of Clearing Fee, Call Proportion in Busy Hours, Call Proportion in Idle Hours, Cost Subsidies and Preferential Fee to the Cost are equal to the normalization of eigenvectors. Consistency test:

$$CI = \frac{\lambda - n}{n-1} = \frac{5.14 - 5}{5-1} = 0.035$$

$$CR = \frac{CI}{RI} = \frac{0.035}{1.12} = 0.03 \leq 0.1$$

If *CR*=0.03< 0.1, it is viewed passing the consistency test. Therefore, weights of Clearing Fee, Call Proportion in Busy Hours, Call Proportion in Idle Hours, Cost Subsidies and Preferential Fee to the Cost are: $(0.355, 0.088, 0.088, 0.268, 0.202)$.

Pairwise comparison matrices of In-the-net Time, Long-distance and Roaming Ratio, Value-added Service Penetration Rate and Data Traffic Saturation to the Satisfaction were constructed:

$$\begin{bmatrix} 1 & 3 & 2 & 3 \\ 1/3 & 1 & 1/2 & 2 \\ 1/2 & 2 & 1 & 2 \\ 1/3 & 1/2 & 1/2 & 1 \end{bmatrix}$$

The eigenvectors corresponding to the largest eigenvalues of the matrices were calculated through the Matlab software: $[0.805, 0.301, 0.465, 0.212]^T$.

Weights of In-the-net Time, Long-distance and Roaming Ratio, Value-added Service Penetration Rate and Data Traffic Saturation to the Satisfaction are equal to the normalization of eigenvectors, $[0.452, 0.169, 0.261, 0.119]^T$.

Consistency test:

$$CI = \frac{\lambda - n}{n-1} = \frac{4.07 - 4}{4-1} = 0.023$$

$$CR = \frac{CI}{RI} = \frac{0.023}{0.90} = 0.026 \leq 0.1$$

If *CR*=0.026< 0.1, it is viewed passing the consistency test. Therefore, weights of In-the-net Time, Long-distance and Roaming Ratio, Value-added Service Penetration Rate and Data Traffic Saturation to the Satisfaction are $(0.452, 0.169, 0.261, 0.119)$.

Finally, weights of indexes to the goal layer could be acquired (Table II).

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.218	32.179	32.179	3.218	32.179	32.179
2	1.989	19.892	52.070	1.989	19.892	52.070
3	1.775	17.747	69.818	1.775	17.747	69.818
4	1.093	10.934	80.751	1.093	10.934	80.751
5	1.000	10.004	90.755	1.000	10.004	90.755
6	.924	9.245	100.000			
7	9.088E-16	9.088E-15	100.000			
8	-2.510E-16	-2.510E-15	100.000			
9	-5.093E-16	-5.093E-15	100.000			
10	-7.617E-16	-7.617E-15	100.000			

Extraction Method: Principal Component Analysis.

Fig. 1. PCA table.

TABLE II: WEIGHTS OF INDEXES

Goal layer	*Criterion layer*	*Weight of criterion layer*	*Index layer*	*Weight of index layer*
Current Value	Profit Contribution	0.670	ARPU	1
	Cost	0.330	Clearing Fee	0.355
			Call Proportion in Busy Hours	0.088
			Call Proportion in Idle Hours	0.088
			Cost Subsidies	0.268
			Preferential Fee	0.202
Potential Value	Satisfaction	1	In-the-net Time	0.452
			Long-distance and Roaming Ratio	0.169
			Value-added Service Penetration Rate	0.261
			Data Traffic Saturation	0.119

In Table II, weights of ARPU, Clearing Fee, Call Proportion in Busy Hours, Call Proportion in Idle Hours, Cost Subsidies and Preferential Fee to the Current Value are 1, 0.355, 0.088, 0.088, 0.268 and 0.202, respectively. Weights of In-the-net Time, Long-distance and Roaming Ratio, Value-added Service Penetration Rate and Data Traffic Saturation to the Potential Value are 0.452, 0.169, 0.261 and 0.119, respectively. Current Value and Potential Value of every customer are the sum of products of weights and corresponding indexes.

2) Establishment of the PCA model

Considering the strong subjectivity of AHP, PCA was employed for data analysis. All data used in this paper are real customer data of a telecom enterprise, including 303 data sizes and 10 attribute values. PCA table is shown in Fig. 1.

Extraction of PCA follows the principle that eigenvalue of the principal component must be higher than 1 and the cumulative contribution rate is 80% at least. Eigenvalue could reflect impact of principal components on factors to a certain extent. When eigenvalue<1, the principal component is not so influential. Generally, only principal components with eigenvalue higher than 1 will be used as consideration factors.

Cumulative contribution rate of five principal components reaches 90.755%. The factor loading matrix shows that ARPU, Preferential Fee and Cost Subsidies possess high loading in the first principal component; Data Traffic Saturation has high loading in the second principal component; In-the-net Time has high loading in the third principal component; Clearing Fee has high loading in the fourth component; Call Proportion in Busy Hours and Long-distance and Roaming Ratio have high loading on the fifth principal component.

Divide each column of the principal component loading matrix by the extraction of the eigenvalue of the principal component is the eigenvector of the eigenvalue, that is, the coefficient of the principal component in each index. Therefore, expression of principal components is:

$$F_1 = 0.4884 \times X_1 + 0.1190 \times X_2 + 0.1602 \times X_3 + 0.0902 \times X_4 + 0.4953 \times X_5 + 0.5095 \times X_6 + 0.2818 \times X_7 + 0.1103 \times X_8 + 0.2939 \times X_9 + 0.1748 \times X_{10}$$

$$F_2 = -0.2019 \times X_1 - 0.0078 \times X_2 + 0.4765 \times X_3 + 0.4893 \times X_4 - 0.1884 \times X_5 - 0.0573 \times X_6 - 0.0384 \times X_7 + 0.0377 \times X_8 - 0.0310 \times X_9 + 0.6713 \times X_{10}$$

$$F_3 = -0.2344 \times X_1 - 0.0158 \times X_2 - 0.1138 \times X_3 + 0.0452 \times X_4 - 0.2214 \times X_5 - 0.2550 \times X_6 + 0.6274 \times X_7 + 0.1369 \times X_8 + 0.6357 \times X_9 - 0.0492 \times X_{10}$$

$$F_4 = -0.1639 \times X_1 + 0.8830 \times X_2 - 0.0665 \times X_3 + 0.0819 \times X_4 + 0.1983 \times X_5 - 0.1746 \times X_6 + 0.0371 \times X_7 - 0.3345 \times X_8 - 0.0211 \times X_9 + 0.0092 \times X_{10}$$

$$F_5 = 0.0653 \times X_1 + 0.1022 \times X_2 - 0.5502 \times X_3 + 0.5475 \times X_4 + 0.1003 \times X_5 - 0.0934 \times X_6 - 0.1714 \times X_7 + 0.5746 \times X_8 - 0.0689 \times X_9 - 0.0127 \times X_{10}$$

where F_1, F_2, F_3, F_4, F_5 are the first, second, third, fourth and fifth principal components. Corresponding factors include ARPU, Clearing Fee, Call Proportion in Busy Hours, Call Proportion in Idle Hours, Cost Subsidies, Preferential Fee, In-the-net Time, Long-distance and Roaming Ratio, Value-added Service Penetration Rate and Data Traffic Saturation are $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$.

Contribution percentage of each principal component in the total contribution is viewed as the weight of each principal component. On this basis, the comprehensive score model of PCA could be calculated. C_n represents contribution of each principal component, where *n*=1, 2, 3, 4, 5.

$$F = \frac{C_1}{C_{\text{all}}} F_1 + \frac{C_2}{C_{\text{all}}} F_2 + \frac{C_3}{C_{\text{all}}} F_3 + \frac{C_4}{C_{\text{all}}} F_4 + \frac{C_5}{C_{\text{all}}} F_5$$

Substitute data and then,

$$F = 0.355 \times F_1 + 0.219 \times F_2 + 0.196 \times F_3 + 0.120 \times F_4 + 0.110 \times F_5$$

Therefore, comprehensive values of customer could be known.

B. Clustering Analysis on Customer Comprehensive Value

PCA gives the comprehensive value of customers rather than current value or potential value.

In this paper, K-means clustering analysis on customer comprehensive values was implemented. Results are shown in Fig. 2.

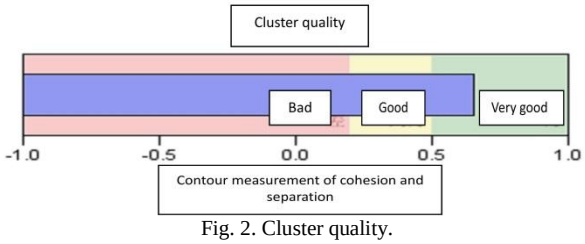

Fig. 2. Cluster quality.

It can be seen from Fig. 3 that K-means clustering could divide customer comprehensive value into four classes: Cluster-1 (28%), Cluster-2 (24%), Cluster-3 (32%) and Cluster-4 (16%).

The following text compares comprehensive values of different data:

Cluster-1, Cluster-2, Cluster-3 and Cluster-4. Cluster-1

represents ordinary customers; Cluster-2 represents customers with very high comprehensive values; Cluster-3 represents customers with high comprehensive values; and Cluster-4 represents customers with low comprehensive values (Fig. 4).

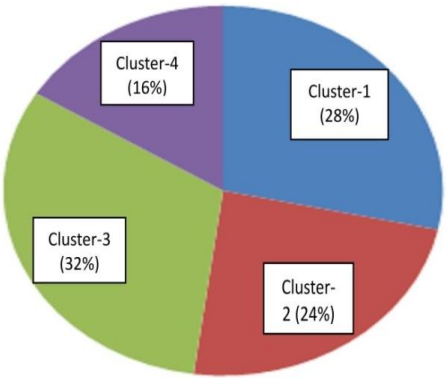

Fig. 3. Cluster distribution.

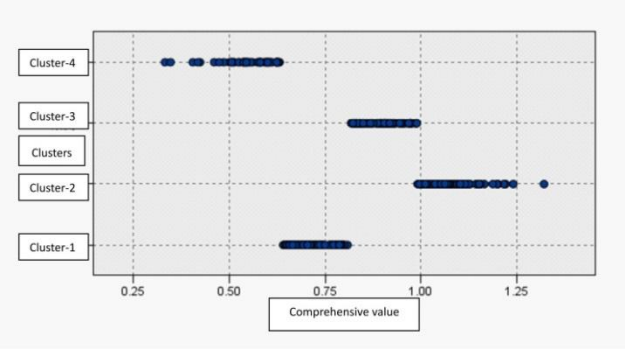

Fig. 4. Comparison of clusters.

III. Conclusions

Customer value determines future development of telecom enterprises. Inadequate resource investment couldn't meet user demands, whereas excessive resource investment will influence enterprise profit and customer value. Telecom enterprises have to transform potential value of customers into enterprise profit and shall adjust resource investment according to customer relations. They shall focus on enhancing customer value as much as possible with limited resources. This paper analyzes customer values through data mining, such as PCA and clustering algorithm. It establishes an evaluation model of customer comprehensive value, classifies customers and finally recognizes different types of customers. Research results could guide telecom enterprises in making differentiated marketing strategies to different customers.

Acknowledgment

The authors would like to thank anonymous reviewers for their constructive and enlightening comments, which improved the manuscript. This work has been supported by grants from Program for Excellent Youth Scholars in Universities of Guangdong Province (Yq2013108). The authors are partly supported by the Key grant Project from Guangdong provincial party committee propaganda department, China (LLYJ1311), Guangdong Natural Science Foundation (NO.2015A030313664, NO.2015A030310340) and Guangzhou science and technology project (NO.201510020013).

References

[1] E. W. T. Ngai, L. Xiu, and D. C. K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2592-2602, 2009.

[2] P. S. Raju, V. R. Bai, and G. K. Chaitanya, "Data mining: Techniques for enhancing customer relationship management in banking and retail industries," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, no. 1, pp. 2650-2657, 2014.

[3] M. Zan, F. Peng, and L. Yanfei, "Customer segmentation of call center IVR based on data mining," *Microcomputer & Its Applications*, vol. 12, 2014.

[4] D. Kang and Y. Park, "Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1041-1050, 2014.

[5] S. H. Liao, P. H. Chu, and P. Y. Hsiao, "Data mining techniques and applications — A decade review from 2000 to 2011," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11303-11311, 2012.

[6] S. Y. Hosseini and A. Z. Bideh, "A data mining approach for segmentation-based importance-performance analysis (SOM–BPNN–IPA): A new framework for developing customer retention strategies," *Service Business*, vol. 8, no. 2, pp. 295-312, 2014.

[7] S. Okazaki, A. M. Díaz-Martín, M. Rozano, and H. D. Menéndez-Benito, "Using Twitter to engage with customers: A data mining approach," *Internet Research*, vol. 25, no. 3, pp. 12-16, 2015.

[8] A. Kazemi, M. E. Babaei, and M. O. M. Javad, "A data mining approach for turning potential customers into real ones in basket purchase analysis," *International Journal of Business Information Systems*, vol. 19, no. 2, pp. 139-158, 2015.

[9] D. Zhang, Y. Ma, X. Tao, and Y. He, "Value analysis of mobile internet users based on clustering," in *Proc. the Ninth International Conference on Management Science and Engineering Management*, Springer Berlin Heidelberg, pp. 447-457, 2015.

[10] Ž. Deljac, M. Randić, and G. Krčelić, "Early detection of network element outages based on customer trouble calls," *Decision Support Systems*, vol. 73, pp. 57-73, 2015.

Xiaoyong Liu is an associate professor who joined the Department of Computer Science, Guangdong Polytechnic Normal University in 2007. He obtained the Ph.D. degree in Graduate University of Chinese Academy of Sciences in 2011. His research interests include text mining, data mining, ant colony optimization and genetic algorithm.

Hui Fu is an associate professor of the Department of Computer Science, Guangdong Polytechnic Normal University. She obtained her master degree from South China University of Technology, China, in 2006. Her major is applied mathematics. Her current research interests include data mining, operations research and genetic algorithm.