An Intergrated Data Mining and Survival Analysis Model for Customer Segmentation

Guozheng Zhang1 Yun Chen2
1College of Business, Houzhou Dianzi University, P.R.China, 310018
(E-mail: guozhengzhang@gmail.com)
2School of Public Economy Administration, Shanghai University of finance & economics, P.R.China, 200433
(E-mail: chenyun@mail.shufe.edu.cn)

Abstract. More and more literatures have researched the application of data mining technology in customer segmentation, and achieved sound effects. One of the key purposes of customer segmentation is customer retention. But the application of single data mining technology mentioned in previous literatures is unable to identify customer churn trend for adopting different actions on customer retention. This paper focus on constructs a integrated data mining and survival analysis model to segment customers into heterogeneous group by their survival probability (churn trend) and help enterprises adopting appropriate actions to retain profitable customers according to each segment's churn trend. This model contains two components. Firstly, using data mining clustering arithmetic cluster customers into heterogeneous clusters according to their survival characters. Secondly, using survival analysis predicting each cluster's survival/hazard function to identify their churn trend and test the validity of clustering for getting the correct customer segmentation. This model proposed by this paper was applied in a dataset from one biggest china telecommunications company. This paper also suggests some propositions for further research.

1 Introduction

Over the past decade, there has been an explosion of interest in customer relationship management (CRM) by both academics and executives (Werner Reinartz, Manfred Krafft, And Wayne D. Hoyer, 2004). Organizations are realizing that customers have

different economic value to the company, and they are subsequently adapting their customer offerings and communications strategy accordingly. Currently research demonstrates that the implementation of CRM activities generates better firm performance when managers focus on maximizing the value of the customer (Gupta, Sunil, Donald R. Lehmann, and Jennifer A. Stuart, 2004). A deeper understanding of customers has validated the value of focusing on them. Customer segmentation is one of the core functions of CRM. Customer segmentation is the base of how to maximize the value of customer (Yun Chen, Guozheng Zhang, et al., 2006). Both researchers and managers need to evaluate and select customer segmentations in order to design and establish different strategies to maximize the value of customers.

Generally, customer segmentation methods mostly include experience description method, traditional statistical methods, and non-statistical methods (Per Vagn Freytag, et al, 2001, Lei-da Chen et al., 2000, E. H. Sub et al., 1999). Non-statistical methods mainly are arisen application of data mining technology in segmentation (Agnes Nairn, and Paul Bottomley, 2003, Verhoef P.C, et al., 2003, Jon Kleinberg, et al., 2004). Jaesoo Kim etc (2003) researched the application of ANN in tour customer segmentation. Fraley, C. and Raftery, A.E (2002) researched the application of clustering approaches in customer segmentation. These literatures use single data mining technology analyses single business issue, and have got some good effectives. But these applications have one obvious big shortages: One of key purpose of customer segmentation is customer retention, previous segmentation methods may be able to position which segment need more care, but it is unable to identify customer churn trend for take effective actions for different customers segments.

The purpose of this article is to propose a new customer segment model.The new customer segment model is called Intergrated Data Mining and Survival Analysis Model for Customer Segmentation. Its effectiveness lies not only in that it identifies key customer segments, but also in that it is able to predict the customer hazard/survival probability (churn trend). Enterprises can make effective customer retention action based on hazard/survival probability (churn trend) of each customer segment. And after segmentation, each customer's hazard/survival probability (churn trend) in one segment is approximately same level, so when calculating customer value, we can use the customer retention rate instead of single customer's hazard/survival probability.

2 Intergrated Data Mining and Survival Analysis Model

2.1 Model Architecture

The intergrated data mining and survival analysis model proposed by this paper contains two key components: Customer Clustering and Churn Trend Identification. They are two indiscerptible parts of the model. Figure 2 shows the architecture of this model.

In Customer Clustering phase, using data mining clustering arithmetic cluster

customers into heterogeneous segments according to their survival characters. And in Churn Trend Identification phase, using survival analysis predicting each segment's survival/hazard function to identify their churn trend and test the validity of clustering. Finally, getting the correct customer segmentation.

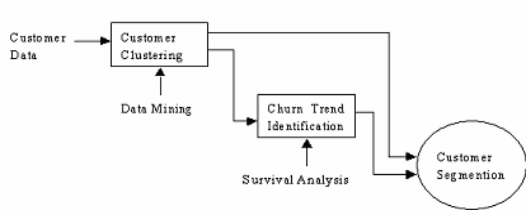

Figure1 Integrated Data Mining and Survival Analysis Model for Customer Segmentation

2.2 Model Implementation Steps

The method consists of cluster technology and survival analysis, as shown in figur1. The customer character for clustering is extracted according to industry and customer behavior attributes. Take the telecommunication industry as example, customer's communication hours, times and expense in recent several months are widely used for churn forecast and they are explainable exactly. So this paper uses these behavior attributes as segmentation variables in empirical analysis.

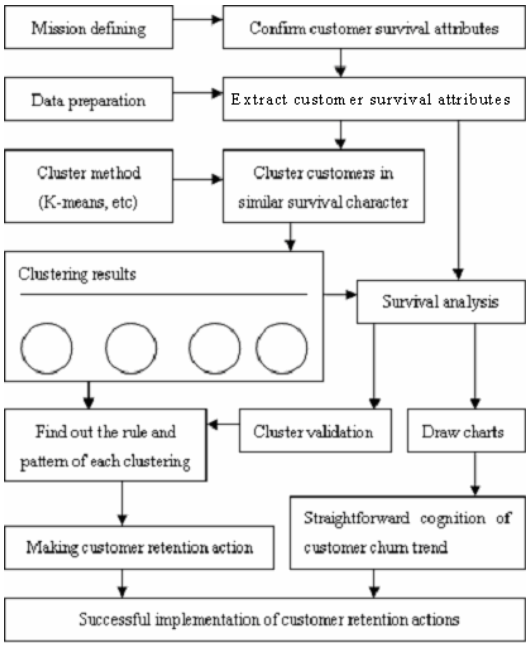

Figure2 Customer Segmentation based on survival character

2.3 Steps

(1) Mission defining. Confirming customer attributes for mission.
(2) Data preparation and Data extraction. Extracting necessary customer behavior attributes from data warehouse.
(3) Clustering. Using data mining clustering methods such as K-means to cluster customers based on similar hazard/survival possibility.
(4) Appending a new attribute: Cluster Number. After clustering, each customer has a new attribute named Cluster number, its necessary for survival analysis.
(5) Survival analysis. It just needs three necessary attributes: Months with Service, Customer Number and Clustering Number. Survival analysis have two purpose: ①Segmentation's quality can be measured by homogeneity in subdivisions and heterogeneity among subdivisions（Wedel M, Kamakura W A, 2000）. It can make comparer between different clusters to test clustering performance and help getting the correct customer segmentation. ②Drawing survival function curve for straightforward cognition of customer churn trend.
(6) Find out the rule and pattern of each clustering.
(7) Successful Implementation of customer retention action

3. Empirical Analysis

A china telecommunications company, which throngs more and more customers, but on the other hand, strives with stronger competition, is our studying case.

3.1 Data selected and filtering

The company wants to make decision to satisfy customers, and prevent customer churn. In collaborate with the company, it supplies with us the research data. Customer data are selected and filtered, and deleted some insignificant records, such as register without transaction records. In the end, we select 1000 records from data warehouse, each record including256 attributes.

Attributes list as follow: customer's basic information(name, gender, register date), churn flag, customer's transaction records from first month to sixth month after register（such as total numbers each month, total fee, the number of calling in and calling out each month, fee in every month, roaming about each month, the number of note, the number of calling in and calling out in working day, the number of transactions in one net and the time, the number of transactions among nets and the time）, the number of customers' consultation, etc..

In the selected dataset, 27.4% customers occur with churn among all customers, namely 72.6% shares with our service. In this paper, we don't list all of attributes, a part of attributes have been listed as showed in table1.

Table 1: customers' behavior attributes

Attributes	Explains	Attributes	Explains
Tenure	Months with service	Total_Times_1	Total number of transaction in first month
Age	Age in years	Total_Times_2	Total number of transaction in second month
Marital	Marital status	Total_Times_3	Total number of transaction in third month
Address	Years at current address	Total_Times_4	Total number of transaction in fourth month
Ed	Level of education	Total_Times_5	Total number of transaction in fifth month
Employ	Years with current employer	Total_Times_6	Total number of transaction in sixth month
Retire	Retired	Total_Duration_1	Total duration in first month
Gender	Gender	Total_Duration_2	Total duration in second month
Reside	Number of people in household	Total_Duration_3	Total duration in third month
Fields Name	description	Total_Duration_4	Total duration in fourth month
Churn Flag	churning flag	Total_Duration_5	Total duration in fifth month
Total_Disc_Fee	Total discount in the past six months	Total_Duration_6	Total duration in sixth month
...			
...			

3.2 clustering customer based on similar survival character

A K-means cluster analysis was performed. According to former literatures with the prediction in customers churn, we selected and filtering 196 attributes to cluster, and omitting some of attributes, like worthless attributes or inapplicable in K-means, especially like sparse datum (Pang-Ning Tan, Michael Steinbach, Vipin Kumar, 2006). Generally, the parameter must be appointed in the beginning with K-means. Now according with advanced experience, the parameter K is set as 2-6. The studying uses the software of SPSS, the result finally arrived at four clustering .Parts of clustering results as showed in table2

Table2 clustering results

Customer ID	Churn flag	Cluster ID	Number of customers in each cluster
1	1	cluster2	
2	1	cluster4	
3	0	cluster4	Cluster1 :194
4	0	cluster2	Cluster2 :264
5	1	cluster3	Cluster3 :269
6	0	cluster2	Cluster4 :273
7	1	cluster4	
8	0	cluster1	
...	...	...	

3.3 Verifying the clustering

The aim to estimate the survival and hazard function help us obtain customer survival/churn information. Necessary variable: Months with service, Churn flag, Customer ID. Firstly, these cluster were been pairwise compared. As showed in table3, their difference is distinct (all sig. <0.10, almost all sig. <0.05). And we get survival function(churn trend) , as shows in Figure3.

Table3 Pairwise Comparisons

Cluster ID	Compared cluster ID	Wilcoxon (Gehan) Statistic	Sig.
1	2	18.640	.000
	3	37.154	.000
	4	2.949	.086
2	1	18.640	.000
	3	5.515	.019
	4	9.222	.002
3	1	37.154	.000
	2	5.515	.019
	4	27.229	.000
4	1	2.949	.086
	2	9.222	.002
	3	27.229	.000

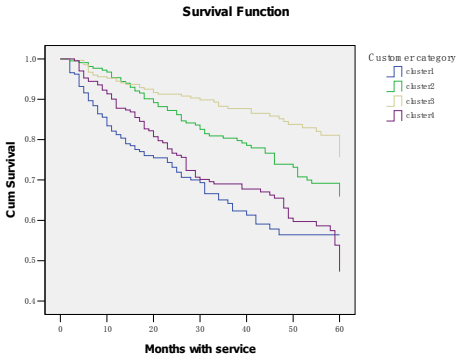

Figure3 survival function

4 Management Applications

With the analysis to all kinds of features, we can find the rule and pattern as showed in Table4, and in the end find relevant market tactics.

Table4 cluster features

class	
1	More fees, more long-distance transactions, highest churn
2	Normal fees, second lowest churn
3	Little transaction each month and more note, lowest churn
4	More transactions each month, more call in and less call out, second highest churn

Clustering1: these customers share important similar features, such as more fees,

more long-distance transactions and highest churn. As these customers expend more, and their expectation is high, we must hold up them with more resource, so as to decrease the churn. And as showed in Figure3, the churn trend is much quicker in the beginning phase that latter phase, so firms must adopt retention actions as soon as possible.

Clustering2: generally, these customers hold average expenditure each month, and every numeric displays equilibration, not higher or lower. The churn is low comparatively. Relatively, these customers satisfy with the company's service, and they hope to share with the company's service. The company must launch into appropriate resource to hold them, for they are the foundation of the customers.

Clustering3: these customers' expenditure is lowest comparatively, and the probability of churn also is also lowest comparatively. They satisfy present services provided with the company, but contribution is lower than average level. Company needn't distribute any resource on them.

Clustering4: these customers have more spending, with more call in and lower call out. They are valuable customers, but their churn is also high. Company must allocate some resources to retain them and encourage them to call out more.

5 Conclusions

A key role of marketing is to identify the customers or segments with the greatest value-creating potential and target them successfully with corresponding marketing strategies to reduce the risk of these high lifetime value customers defecting to competitors (Andrew Banasiewicz, 2004). Segmenting customer is the basic work of data mining according to known historic segmentation information. The training data used to construct segment forecast mode can be historic data or exogenous data that gain from experience or survey.

For an enterprise, how to use data mining, and how to practice enterprise's tactics should we use in determine segmentation? To answer this question, this paper proposes a Integrated Data Mining and Survival Analysis Model for Customer Segmentation. The model proposed by this paper has expanded the simple application of single data mining technology in customer segmentation, it can guide for more complex applications of data mining in CRM. The model clusters customers into heterogeneous clusters with similar churn probability, each clustering have its unique churn trend, so enterprise can make retention action according to this. And via observe the sharp of survival function curl, enterprises can get straightforward cognition of customer churn trend.

Effective segmentation can help companies increase revenue by acquiring and retaining high value customer at low cost. It can also help in aligning cost-to-serve to customer value, perhaps reducing overall marketing, sales and service costs. The model proposed by this paper is testing in the telecomm industry; it may be used in other industry such as finance service etc. Therefore, the future researches may focus on testing this model in other service industry.

Acknowledgements

The Research is supported by the Young Teacher Research Start-up Project "Research of Customer Lifetime Value Management System Basded on Data Mining Technology" of Houzhou Dianzi University under the grant No: Y021507037.

References

26. 1. A. Nairn, and P. Bottomley (2003), Cluster analysis procedures in the CRM era, *International Journal of Market Research*, Vol. 45 Quarter 2
27. 2. A. Banasiewicz, Acquiring high value, retainable customers, *Database Marketing & Customer Strategy Management*, 2004, Vol. 12, 1, 21–31
28. 3. E. H. Sub, K .C. Noh, C .K.Suh (1999), Customer list segmentation using the combined response model, *Expert Systems with Applications*, 17(2): 89-97
29. 4. C. Fraley, and A.E. Raftery(2002), *Model-Based Clustering, Discriminant Analysis, and Density Estimation.* Journal of the American Statistical Association 97:611-631.
30. 5. G. Sunil, D.R. Lehmann, and A.S. Jennifer (2004): *Valuing Customer Journal of Marketing Research*, , 41 (February), 7–18.
31. 6. JR. DW. Hosmer, and S. Lemeshow(1999), *Applied Survival Analysis: Regression Modeling of Time to Event Data*, New York: John Wiley & Sons,.
32. 7. K. Jaesoo et al.(2003): Segmenting the market of West Australian senior tourists using an artificial neural network, *Tourism Management*, 24(l):25-34
33. 8. J. Kleinberg, C. Papadimitriou, P. Raghavan (2004): *Segmentation Problems, Journal of the ACM*, Vol. 51, No. 2, March, pp. 263–280.
34. 9. L.D. Chen, K.S. Soliman, E. Mao, M. N. Frolick (2000): Measuring user satisfaction with data warehouse: an exploratory study, *Information & Management*, 37(3): 103--110.
35. 10. P.N. Tan, M. Steinbach, V. Kumar (2006). *Introduction to Data Mining*, Posts & Telecom Press, pp310-320, Beijing.
36. 11. P.V. Freytag, et al(2001), Business to Business Market Segmentation Industrial Marketing Management，30(6):4 73486
37. 12. Y. Chen, G.Z. Zhang, D.F Hu and S.S. Wang (2006), Customer Segmentation in Customer Relationship Management Based on Data Mining, *Knowledge Enterprise: Intelligent Strategies in Product Design, Manufacturing, and Management,* Volume 207, pp.288-293
38. 13. P.C. Verhoef, P.N. Spring, J.C. Hoekstra , The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands, *Decision Support Systems*, 2 003,34(4):471-481
39. 14. M. Wedel, W.A. Kamakura and U. Bockenholt (2000): Marketing data, models and decisions, *International Journal of Research in Marketing* 17(2-3) 203-208.
40. 15. W. Reinartz, M. Krafft, and W.D. Hoyer(2004), The Customer Relationship Management Process: Its Measurement and Impact on Performance, *Journal of Marketing Research*, 293 Vol. XLI (August), 293–305）