Clustering Analysis for Silent Telecom Customers Based on K-means++

Yuhang Qiu[1], Pingping Chen[1], Zhijian Lin[1], Yongcheng Yang[2], Lanning Zeng[1], Yaqi Fan[1]
1. Department of Electronic Information Engineering, Fuzhou University, Fujian, China
2. Department of Navigation, Jimei University, Fujian, China
qyu378295929@163.com, ppchen.xm@gmail.com, zlin@fzu.edu.cn, lanningz@outlook.com, 18973485719@sina.cn

Abstract— **Silent customers are part of customers that company is very easy to lose. It is necessary to analyze the features of such customers and make appropriate market decisions to improve the enterprise's revenue in telecom industry. This paper proposes a K-means++ method for customer segmentation based on silent customers. Firstly, key variables to the segmentation model was screened out and then the original data was preprocessed. Secondly, silent customers were clustered and the Calinski-Harabasz index was adopted to verify the best clustering effect when k=6. At last, radar chart analysis and suggestions were given, which would provide data supports to the improvement of operation and maintenance management and decision-making of the precision marketing.**

Keywords—Silent customer; Customer segmentation; Telecom industry; Clustering; K-means++

I. Introduction

Generally, customers are the most important asset to the revenue of business companies. With the rapid development of computer technologies, more and more companies are building up their databases of customers and using machine learning algorithm to analyze large customer data, like clustering[1,2]. The process of discovering unknown and potentially useful information and knowledge from large amounts of data is called data mining, which is able to help enterprises to determine the market position and make proper marketing decision. Researches on data mining in the company have been done by some previous researchers, such as customer relationship management of banking[3], customer analysis in the electric vehicle industry[4].

Telecom industry is a typical data-intensive industry, and the communication customers will continuously produce a large amount of data. For these data, one of the important applications of data mining in telecom industry is customer segmentation[5]. Customer segmentation is to classify customers into different groups according to their attributes. The customers within the same group have greatest similarity, and the ones from different group have greatest difference. Customer segmentation can achieve following objectives: (1) to understand the customer's composition; (2) to understand the group characteristics of various customers; (3) to identify potential lost customers. In fact, customer segmentation problems can be transformed into clustering problems. And K-means [6] is a classical algorithm to solve the clustering problem. Due to the high efficiency of this algorithm, many researchers have used it to carry out relevant researches in different tasks[7][8].

However, most researchers only focus on the data of current customers and analyze their characteristics to develop strategies to attract new customers. Actually, the cost to the company of acquiring potential customers is much higher than the cost of saving lost customers[9]. Therefore, it is easier for companies to increase revenue at low cost by analyzing the customer data that has been lost or being lost and make some marketing decisions. On the other hand, traditional K-means algorithm has an obvious defect which is to determine the clustering center by random selection. It will lead to different clustering results. In order to solve this problem, it reasonable for us to adopt K-means++ algorithm[10] which is an improved algorithm of K-means to make different clustering centers as far away as possible so as to accurately segment customers.

This paper performed extraction of silent customer dataset in a large communication company of Fujian Province. We segment silent customer using K-means++ clustering method. The results show the clustering situation of silent customer data and radar charts containing various characteristic trends. At the same time, we also give suggestions to the company's market decisions according to clustering results of different customers.

The remainder of this paper is organized as follows. Section II provides the proposed business problem, the introduction of dataset, method structure, data preprocessing and related algorithm. Section III shows the results of clustering and suggestions. Finally, Section IV concludes this paper briefly.

II. Material and Analytical Method

A. Definition of business problem

This paper takes the silent customers of a large communication company in Fujian Province as a case to study. Silent customers in communication industry refer to those who are active at a previous time and have frequent communication session, but are less active at the current time and rarely communicate. These are the customers that have lost or are losing. Using data mining technique of clustering analysis, we can group the silent customer, characterize each group and analyze their properties to adopt different strategies for these customers, so as to save and avoid the loss of silent customers.

B. Dataset

We collected 125,296 silent customers data from a large communication company in Fujian Province as our dataset. This dataset includes all silent customers of this company in a city

978-1-7281-4390-3/20/$31.00 ©2020 IEEE

within certain months. We obtained data of silent customers in October and November this year to ensure timeliness of analysis results. The original data of silent customer contains 26 features, which can be divided into the following categories: customer's identity information, customer's package fee, package usage and customer's communication frequency.

C. *Proposed Method*

The research design consists of several stages including data collecting, data preprocessing, clustering with K-Means++ algorithm and suggestion for market decision. The proposed method based on Grid Hill economic theory approach[11]. The flow chart of method is shown in Figure 1.

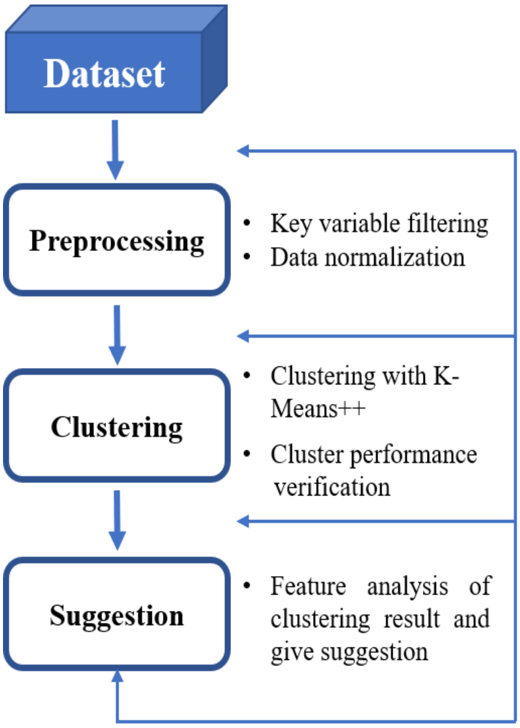
Dataset
Preprocessing
• Key variable filtering
• Data normalization
Clustering
• Clustering with K-Means++
• Cluster performance verification
Suggestion
• Feature analysis of clustering result and give suggestion

Figure 1. The flow chart of proposed method.

D. *Key variables filtering.*

Aiming at determining input variables of customer segmentation model, we referred to the classical RFM model for customer value assessment and analyzed the correlation of 26 variables. Based on the requirements of the company and the suggestions of business experts, 6 uncorrelated variables were screened out finally as input variables for the customer segmentation model, which are:

TABLE I. KEY VARIABLES FILTERING RESULT

Name	**Abbreviation**	***Description***
Age	A	The age of customer
Net age	N	Duration as a customer of the company
Communication days	C	The number of communication days in given months
Package fee	P	The cost of the package purchased by customers
Balance	B	The balance of the customer's account
The number of base stations passed	BS	Number of communication switches between different locations

In those selected variables, age and net age stand for characteristic information of customers. They are necessary in the communication industry, where companies often come up with a variety of packages based on age groups. For communication days, package fee, balance and the number of base stations passed, we only analyze the data in October and November. Communication days and the number of base stations passed reflect the customer's activity. The package price actually represents the company's revenue, while the balance can reflect the possibility of the customer's subsequent reactivation. The higher the remaining balance, the customer tend to spend it and become active again. Table II shows the partial data information of several key variables after filtering.

TABLE II. DATASET OF CUSTOMER

Use ID	**A**	**N**	**C**	**P**	**B**	**BS**
C01	59	37	0	38	50	0
C02	48	32	8	38	100	53
C03	45	27	5	18	47.2	4
C04	26	8	0	18	19.37	0
C05	33	18	17	38	34	11
…	…	…	…	…	…	…
C125296	30	4	2	18	12	4

E. Data normalization

In order to avoid the influence of input variables with different unit dimensions in clustering model on distance calculation, it is necessary for us to standardize the key variables. After filtering some abnormal data, we adopted zero-mean normalization and standardized formula which can be given by:

$$x' = \frac{x - x_{mean}}{s} \tag{1}$$

$$x_{mean} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - x_{mean})^2} \tag{3}$$

where, x' is the standardized value, n is the number of data, x_{mean} and s represent the mean and variance of selected variables respectively.

F. K-mean++ algorithm

The random selection of the initial cluster center makes the traditional K-means converge to the local optimal solution easily, it fails to deliver the best clustering results for business customers, making corporate decisions wrong. This is not a good choice for a commercial dataset analysis. In view of the shortcomings of the K-means algorithm, the K-means++ algorithm was adopted in our research. The initial point of K-means++ were randomly selected and make the distance of different center points as far as possible, so as to get the global optimal result. The K-means++ algorithm outperforms traditional K-means algorithm, both in terms of stability and accuracy of clustering result. The specific algorithm process of K-means++ is shown in Figure 2.

```
Input: A set of data points $X$; the number of $k$ clusters; the distance of point $dp$
Output: $X$ points are grouped into $k$ clusters respectively
Set $m$ be the number of points in $X$
1.Select a center $c_1$ randomly from points $X$.
2.For $i$ in 1.. $k$-1 do:
    For j in 1..m do:
        $d_{i,j} = dp(c_i, x_j)$
        $mds_j = [min_{1\le y\le i}(d_{y,j})]^2$
    End for
    Choose next cluster center $c_{i+1}$ from $x_j \in X$ with probability
        $p(x_j) = \frac{mds_j}{\sum_{p=1}^{m} mds_p}$
3.End for
```

Figure 2. The flow chart of K-means++ algorithm.

G. Calinski-Harabasz index evaluation

The process of evaluating the effectiveness of a cluster is called cluster validation[13]. Since our unsupervised clustering has no sample output, there is no direct clustering evaluation method. However, we can evaluate the effect of clustering from the density degree with clusters and the dispersion degree between clusters. Calinski-Harabasz index belongs to this group of methods. It is expressed as a ratio of between-cluster variance and the overall within-cluster variance:

$$I_{CH} = \frac{tr(B_k)}{tr(W_k)}\frac{m-k}{k-1} \tag{4}$$

where m and k are the number of training samples and categories respectively. B_k is the covariance matrix between categories, and W_k is the covariance matrix for the data of categories. *Tr* is the trace of matrix. In other word, the smaller the covariance of the data within category, the greater the covariance between the categories, the higher the Calinski-Harabasz index will be and the better clustering effect will be. In a previous comparative study this index was demonstrated to be one of the best cluster validation tools[14], it is therefore used in this paper as a crucial evaluation to test the performance of clustering results.

III. Simulation Results and Analysis

A. Clustering result

Based on K-means++ algorithm, we carried out experiments repeatedly and finally found that the clustering result effect was the best when k=6. The algorithm went through a total of 16 iterations to stabilize. The customers were divided into 6 clusters, with clustering results being shown in Table III, where the observation of cluster 1 means is (-0.59, -0.26, -0.33, -0.11, -0.17, -0.19) for instance. And the number of different clusters can be seen in Table IV.

TABLE III. The Means of Cluster

Cluster	A	N	C	P	B	BS
1	-0.59	-0.26	-0.33	-0.11	-0.17	-0.19
2	1.24	-0.28	-0.29	-0.20	-0.18	-0.19
3	-0.32	0.43	3.13	2.15	0.42	7.03
4	-0.19	-0.10	2.35	0.67	0.03	0.79
5	0.22	3.00	0.10	0.19	0.43	-0.12
6	0.02	0.84	-0.06	0.51	5.66	0.44

TABLE IV. THE NUMBER OF DIFFERENT CLUSTERS

No	Cluster	Number of data
1	Customer 1	62603
2	Customer 2	30193
3	Customer 3	1407
4	Customer 4	10975
5	Customer 5	7894
6	Customer 6	2078

B. *Clustering performance verification*

A cluster will become convergent if there is no change or movement of members from one cluster to another. From the clustering result, we used Calinski-Harabasz index to evaluate our clustering result. Of the six test scenarios that have been performed, a ranking will be made based on Calinski-Harabasz index and is shown in Figure 3.

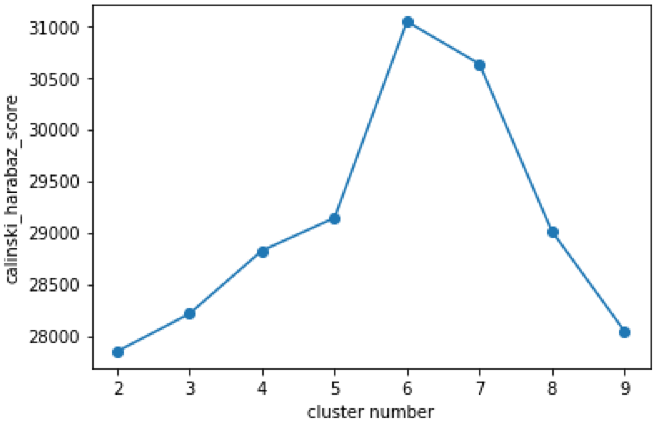

Figure 3. The Calinski-Harabasz value for various numbers of clusters.

The result shows that the number of cluster 6 has the greatest Calinski-Harabasz value among the other clusters of 31,039 listed in Figure 3. It proves the correctness and validity of choosing k=6 in our previous clustering process.

C. *Radar chart*

In order to reflect the feature distribution of different categories after clustering and comparing the difference of them, we used radar chart to display the result and it is shown in Figure 4. The center part of the circle represents the curve distribution of six features, and the higher value of feature, the larger the curve diffusion. The upper right corner of radar chart shows the different categories of clustering results. Six different categories are represented by six different colors and correspond to the curves in the circle.

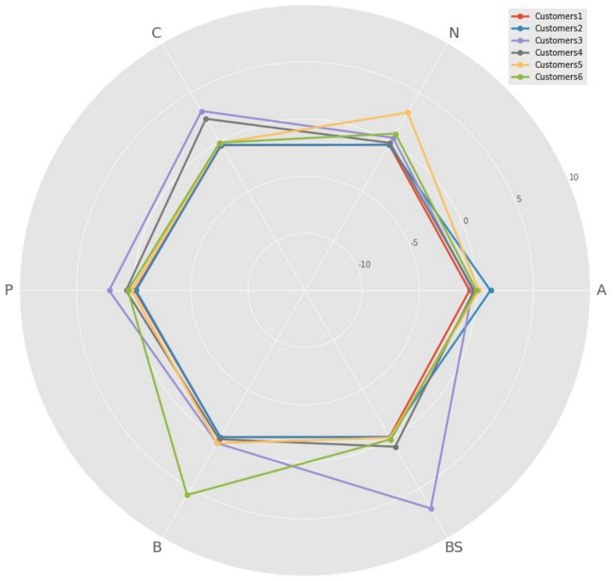

Figure 4. The radar chart of clustering results.

From the result of radar chart, we can clearly find the cluster of Customer 1 and Customer 2 to contain the most data, but the value of all their features except identity information are very low. Such users can be judged to be completely silent. Customer 3 and Customer 4 are relatively active, but Customer 4 brings less revenue to the enterprises. And Customer 6 has the most balance.

D. *Suggestion for company's market decision*

According to the result of cluster clustering and radar chart, we analyzed customers and made some recommendations. Table V shows the relevant content.

TABLE V. CUSTOMER ANALYSIS AND SUGGESTION

Customer Type	Cluster	Customer Characteristics
Important activable customers	Customer 3,4	Mainly for the young group who just become the company's customers. They have high communication frequency and can bring certain income to the company.
Ordinary activable customers	Customer 5,6	Mainly for the middle-aged group, customer 5 is the company's old customers, easy to have feelings for the company. And the customer 6's account balance is large, the customer easily wants to spend it and active again.
Completely silent customer	Customer 1,2	They have the largest number of people and they are mainly composed of teenagers and the elderly. Both the communication frequency and profit contribution are very low and hard to activate.

IV. CONCLUSION

In this work, we adopted K-means++ clustering algorithm to segment silent customers of a large communication company. With the help of Calinski-Harabasz index, we prove the effectiveness of our clustering results. After that, we used radar chart to compare the features of customers in different clusters. Finally, we analyzed the types of silent customers and proposed suggestions for market decisions. The overall process can be concluded as follows: data collection, data preprocessing, clustering and suggestion.

REFERENCES

[1] J. Zhao, W. Zhang and Y. Liu, "Improved K-Means cluster algorithm in telecommunications enterprises customer segmentation," *2010 IEEE International Conference on Information Theory and Information Security*, Beijing, 2010, pp. 167-169.

[2] N. R. Maulina, I. Surjandari and A. M. M. Rus, "Data Mining Approach for Customer Segmentation in B2B Settings using Centroid-Based Clustering," *2019 16th International Conference on Service Systems and Service Management (ICSSSM)*, Shenzhen, China, 2019, pp. 1-6

[3] Jayasree, "A review on data mining in banking sector[J]," American Journal of Applied Sciences, 2013, 10(10):1160-1165.

[4] L. Zhang, P. Wang, P. Chen, X. Li, B. Zhang and L. Ma, "Customer Segmentation Algorithm Based on Data Mining for Electric Vehicles," *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, Chengdu, China, 2019, pp. 77-83.

[5] S. Donner, "What can customer segmentation accomplish[J]," Bankers Magazine, 1992, 175(3/4):72-81.

[6] L. He, L. Wu and Y. Cai ,"Survey of Clustering Algorithms in Data Mining[J]" Application Research of Computers 2007, 24(1):10-13.

[7] L. Julia, B. Julien and C. Filippo, "K-Means Clustering for Data Visualization and Flow Interpretation: Inclined Jet in Crossflow Example." *Aps Division of Fluid Dynamics Meeting* APS Division of Fluid Dynamics Meeting Abstracts, 2013.

[8] P. Tang, W. Qiu, M. Yan, Z. Huang, S. Chen and H. Lian, "Association Analysis of Abnormal Behavior of Electronic Invoice Based on K-Means and Skip-Gram," *2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC)*, Hangzhou, China, 2019, pp. 300-305.

[9] B.A. Tama, "Penetapan Strategi Penjualan Menggunakan Association Rules dalam Konteks CRM[J]." Jurnal Generic, 2013.

[10] A. David , and S. Vassilvitskii, "K-Means++: The Advantages of Careful Seeding." *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007* ACM, 2007.

[11] D. Carole, "Handbook of Customer Satisfaction Measurement[J]." *International Journal of Operations & Production Management* 19.1(1999):95-96.

[12] G. Seni and J. F. Elder, “Ensemble methods in data mining: Improving accuracy through combining predictions,”Synthesis Lectures on Data Mining and Knowledge Discovery,vol. 2, no. 1, pp. 1–126, 2010. [Online]. Available:https://doi.org/10.2200/S00240ED1V01Y200912DMK002

[13] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques[J],” Journal of Intelligent Information Systems, vol. 17, no. 2-3, pp. 107–145, 2001

[14] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, “An extensive comparative study of cluster validity indices,” Pattern Recognition, vol. 46, no. 1, pp. 243 – 256, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S003132031200338X