A New Customer Segmentation Framework Based on Biclustering Analysis

Xiaohui Hu[1]
1 Laboratory of Quantum Engineering and Quantum Materials, School of Physics and Tele-communication Engineering , South China Normal University, Guangzhou 510006, China
Email: xiaohui_huhu@sina.com

Haolan Zhang[2] , Xiaosheng Wu[1], Jianlin Chen[1],Yu Xiao[1], Yun Xue[1] , Tiechen Li[1], Hongya Zhao[3]
2 NIT, ZheJiang University, Hanzhou, P.R.China
3 Industrial Center, Shenzhen Polytechnic, Shenzhen , Guangdong, China
Email:haolan.zhang@gmail.com, super-fly@foxmail.com

***Abstract*—The paper presents a novel approach for customer segmentation which is the basic issue for an effective CRM (Customer Relationship Management). Firstly, the chi-square statistical analysis is applied to choose set of attributes and K-means algorithm is employed to quantize the value of each attribute. Then DBSCAN algorithm based on density is introduced to classify the customers into three groups (the first, the second and the third class). Finally biclustering based on improved Apriori algorithm is used in the three groups to obtain more detailed information. Experimental results on the dataset of an airline company show that the biclustering could segment the customers more accurately and meticulously. Compared with the traditional customer segmentation method, the framework described is more efficient on the dataset.**

***Index Terms*—Customer segmentation, biclustering, K-means,Chi-square statistics, DBSCAN, Apriori**

I. INTRODUCTION

In today's highly competitive business environment, customer relationship management (CRM) is a critical success factor for the survival and growth of businesses, which has been more widely used in some industries and areas, including tourism, catering, retail trade, network marketing, network services and other e-commerce, etc. Customer segmentation is the basic issue for an effective CRM due to its role in helping organizations to understand and serve existing customers better, and enabling the acquisition of profitable customers. Nowadays, data mining technology plays a more important role in the demands of analyzing and utilizing the large scale information gathered from customers.

Many studies in the literature have researched the application of data mining technology in customer segmentation, and achieved sound effectives. Alex. Berson used decision trees and clustering technology for customer segmentation [1]. Guillem Lefait presented a data mining architecture based on clustering techniques to help experts to segment customer based on their purchase behaviors[2]. Jaesoo Kim used neural networks in tourism industry customer classification[3], Meng Xiaolian, Yang Yu proposed a customer identifying model based on customer value in commercial banks[4].

Literatures [6][7] select the K-means clustering algorithm to recognize groups of customers who share the same or similar needs [5]. The K-means clustering algorithm has been widely used because of its simplicity and its efficiency. Segmentation is done not only to identify groups of entities that have common characteristic but also to better understand consumer behaviors. However, when we use clustering to segment customers there exist some problems: which data to select, how many clusters to produce and how to evaluate the clustering results [20]. Thus, the customer segmentation has two main challenges. The first challenge is to formalize implicit data and the second challenge is to select a relevant subset of the available features to perform the clustering.

Hence there exist limitations in clustering algorithms such as K-means or hierarchical clustering, etc. Firstly, common cluster methods usually seek a disjoint cover of the set of elements, requiring that no objects belong to more than one cluster. In fact, customers can participate in more than one activity and should therefore be included in several clusters. What's more, clustering algorithms obtain clusters in either rows or columns. In this case, the clusters produced reflect the global patterns of data. Therefore clustering fails to detect local patterns in the data. Biclustering, or subspace clustering, was proposed to overcome above problems of traditional clustering. Biclustering performs simultaneous clustering on the rows and columns of the data matrix, which is able to find local patterns in the form of subgroups of rows and columns.

The paper proposes a new framework for customer segmentation based on biclustering analysis, which could not only formalize the implicit data but also acquired a subset of the available attributes to provide more explicit informations about customers. Firstly, the chi-square statistical analysis is applied to choose set of attributes and K-means algorithm is employed to quantize the value of each attribute. Then DBSCAN algorithm based on

doi:10.4304/jsw.9.6.1359-1366

density is introduced to divide the customers into three groups (the first, the second and the third class). Finally biclustering based on improved Apriori compared with the original Apriori algorithm is used in the three groups to obtain more detailed information.

The rest of this paper is organized as follows. The proposed method is presented in Section II. The results of the experiment are then presented and analyzed in section III. Finally, the conclusion is made in section VI.

II. Customer Segmentation Architecture

Our objective is to produce detailed diverse and meaningful customer segmentations. These segmentations will be used to help experts to discover specificities in consumer behaviors. At first the raw data is preprocessed as following: Designate an attribute according to the business strategy, then analyze the correlation between the designated attribute and the other attributes, after that delete weak correlation and redundancy attributes. Therefor the rest attributes are independent to each other.

As far as the value of the attributes is concerned, K-means algorithm is applied to each attribute and each value is quantized according to its cluster. Secondly the customers are clustered by the DBSCAN algorithm, which are divided into three groups: the first class, the second class, the third class. Traditional customer segmentation methods usually divide the customers into different groups at this step to get the final conclusion, while we will continue based on it to obtain more detailed information.

At the third step, biclustering algorithm based on improved Apriori is introduced for behavior feature clustering, which is compared with the algorithm based on Apriori. Finally, the final results are analyzed and the corresponding marketing strategy is put forward which could provide more meticulous information of the customers than the traditional methods.

The architecture of the system is shown in Figure 1.

A. Data-Preprocess: Chi-squared Statistical Analysis & K-means Quantization

1 Chi-squared statistical analysis

In this paper we utilize the correlation of chi-square statistic[20] to measure the relevance between the arranged attribute and the rest attribute. The attributes are categorized in three levels: the strong correlation, correlation, and weak correlation. Remove redundancy attributes and attributes with the weak correlation, so the rest attributes are independent to each other.

The algorithm is stated as following:

1) Calculated the Chi-value K_i between the designated attribute C and the other attribute F_i.

2) Rearrange each attribute according to the results of 1.

3) Divide all the attributes into three categories, namely the strong correlation (Strong Relevant) subsets, the relevant subsets, the weak correlation subsets. Determine the Weakest-of-Strong-Relevant value FWoS and the Strongest-of-Weak-Relevant value FSoW.

4) Select the reference attribute F_i from the subset of strong correlation (SSR), iterate the other attributes F_j of the subset and calculate the Chi value K_{ij} between them. If K_{ij} is greater than or equal to FWoS, then delete the attribute F_j, otherwise retain it. Put F_i into the subset of Reduction-of-Strong-Relevant (SRSR). If the SSR is not null, continue.

5) Calculate K_{ij} between the attribute F_i of the correlation subset and the attribute F_j of the Reduction-of-Strong-Relevant subset SRSR. If K_{ij} is greater than FSoW, then remove F_i; otherwise retain it and put it into the subset of Reduction- of-Relevant(SRR).

6) Delete the subset of the weakest correlation.

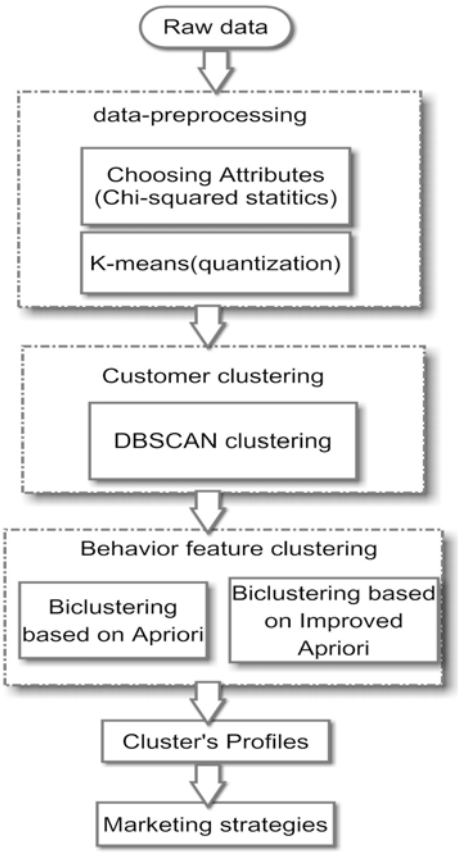

Figure 1. The architecture of customer segmentation

2. K-Means algorithm

K-Means algorithm[11] is a classical algorithm to solve the clustering problem. The idea of specific algorithm is that the k sample points selected randomly are taken as the center of initialized cluster and then performing iterative operations. Clustering results are affected by the choice of initial point, and therefore the solutions obtained are always local optimum, not global optimum.

K-means algorithm is used to cluster the values of each attribute. The value is quantized according to its cluster, which means if it belongs to cluster 1 then its value is changed to 1 and if it belongs to cluster 2 then its value is 2, and so on. The Quantization procedure is shown as Figure2.

MemberID	A	B	C	D
M_1	a_1	b_1	c_1	d_1
M_2	a_2	b_2	c_2	d_2
...	...	...	...	..
M_n	a_n	b_n	c_n	d_n

MemberID	A	B	C	D
M_1	1	2	3	1
M_2	2	1	2	2
...	...	...	...	..
M_n	3	2	2	1

Figure 2. Quantization of the values of attributes

3 Pseudo code

According to the algorithm, the pseudo code in Matlab platform is given below. The original data of each attribute was gradient quantized in order to reduce the resources occupied by the construction of contingency table.

```
INPUT：S（F0,F1，F2...FN）// Gradient quantized attributes subset,C // Classified attribute(Total Price)
OUTPUT：S' // Related subset
a) For i=1:N
T=Table(Fi,C); // Construct contingency table from Fi and C
Ki=Calculate(T); //Caculate K i with Chi-value
b) Sort(S) // Sort Ki ascendingly
c) Rank(S) // Clustered S through K-means algorithm,
output SSR(Strong Relevant Subset)，SR（Relevant Subset）and SWR(Weak Relevant Subset)
d) F=GetFirst(SSR)
while(Fp≠ NULL) // Remove the redundancy attributes in the strong relevant subset
{ Fq= GetNext(SSR, Fp); //Get next one in the subset
while(Fq≠ NULL)
{ tp,q= Table(Fp, Fq);
Kp,q= Calculate(tp,q);
Fq' = Fq;
if(Kp,q≥WSδo)
Delete(SSR, Fq); //Remove the current attribute
Fq= GetNext(SSR, Fq'); }
Fp=GetNext(SSR, p); }
e) F=GetFirst(SR)
while(Fp≠ NULL)
{ Fp' = Fp;
Fq= GetFirst(SSR);
while(Fq≠ NULL)
{ tp,q= Table(Fp, Fq);
Kp,q= Calculate(tp,q);
if(Kp,q>SWδo)
{ Delete(SR, Fp);
break;}
Fq= GetNext(SSR, Fq);
}
Fp= GetNext(SR, Fp'); }
f) S'=Combine(SSR,SR);
g) Return S';
```

B. DBSCAN

Density-based clustering algorithms find groups or regions with high densities which are separated by low density regions [12]. The density based spatial clustering of applications with noise (DBSCAN) algorithm classifies all available points as core points, border points, and noise points. Core points are those that have at least Minpts number of points in the Eps distance. Border points can be defined as points that are not core points, but are the neighbors of core points. Noise points are those that are neither core points nor border points. It has two parameters, a distance parameter Eps and a threshold MinPts. DBSCAN searches for clusters by checking the Eps of each point in the database. If the Eps of a point P contains more than MinPts, a new cluster with P as core object is created. DBSCAN then iteratively collects directly density-reachable objects from these core objects, which may involve the merge of a few density-reachable clusters. The process terminates when no new points can be added to any cluster.

1 K-dist graph

The basic approach of how to determine the parameters Eps and MinPts is to look at the behavior of the distance kth from a point to its k nearest neighbor, which is called k-dist. The k-dists are computed for all the data points for some k, sorted in ascending order, and then plotted using the sorted values, as a result, a sharp change is expected to see. The sharp change at the value of k-dist corresponds to a suitable value of Eps.[21][22]

However, it spends a lot of time and consumes resources on searching and drawing K-dist graph. In order to realize this task better, we use the parallel computation based on Fork/Join Java framework. The total task is divided into multiple sub tasks, and the combination of multiple sub tasks results to reach the final answer. The process is stated as Figure3:

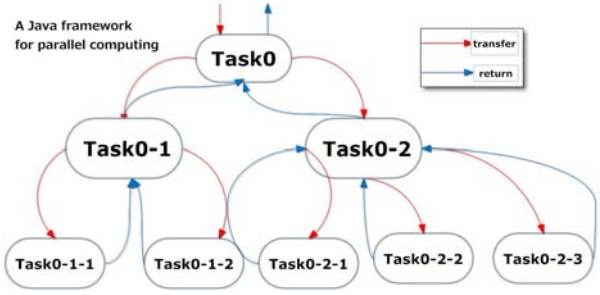

Figure 3. The process of the parallel computation based on Fork/Join Java framework

2 The DBSCAN algorithm

The DBSCAN algorithm is given as following [12].

1) Begin.

2) Arbitrary select a point p.

3) Retrieve all points density-reachable from p, set Eps and MinPts.

4) If p is a core point, a cluster is formed.

5) If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.

6) Continue the process until all of the points have been processed.

7) End.

In this paper, the customers are divided into three groups (the first, the second and the third) through the algorithm.

C. Biclustering Algorithm

Bicluster is a hot topic in data mining, which plays an important role in applications of the gene expression data. A bicluster is a submatrix in the given data matrix, which has certain consistency between its elements. However, bi-clustering clusters both rows and columns which is a NP hard problem. Besides traditional CC (Cheng and Church algorithm)[14] algorithm, there are few mature algorithms to solve it. We change the biclustering problem completely into mining the frequent patterns based on the association rules, so the mature algorithms in association rules can be used for biclustering.

Association rules mining seeks interrelations hidden between data entries, its general objects are transaction databases. Hence we transform the original database into the transaction database after the first two steps.(Figure4.).

MemberID	A	B	C	D
M_1	1	2	3	1
M_2	2	1	2	2
...	...	...	...	..
M_n	3	2	2	1

MemberID	Transactions
M_1	A1,B2,C3,D1
M_2	A2,B2,C2,D2
...	...
M_n	A3,B2,C2,D1

Figure 4. The transformation of the original database into the transaction database

1 Apriori

Apriori is one of the most classical algorithms in mining association rules whose core component is frequent pattern mining. The first step of the algorithm is to identify all the frequent item sets with the condition that their support degrees are greater than the minimum support degree, then searching for (K+1) item sets through the K item sets by layered search iterative method based on frequency set theory. The last step is to form rules form the frequent items guarantee the confidence not less than the minimum confidence user given[21] thus the relationship between the database projects could be found.

The algorithm introduces heuristic thought based on the following: all non-empty frequent item set must also be frequent, if an item set is not frequent, then any of its supersets are non-frequent. The frequent pattern connection pruning step greatly reduces the candidate sets generated, thereby improves the efficiency of the frequent pattern mining. However, it will be very costly if there exist a large number of frequent patterns or very long patterns or minimum support threshold is small. At the same time, it also requires a lot of I/O loads for repeatedly scanning the transaction database.

2 The improved Apriori based on TBBMI

To avoid the weakness of Apriori, the prefix tree, a special kind of data structure, is introduced to store and traverse the frequent itemsets. The prefix tree supports easy processing of database and allows us to store itemsets efficiently by using less memory[24]. It only needs two times to scan the original database. The divide and conquer strategy is employed without generation of candidate itemsets[22][23].

A new structure called TBBMI (Tree Based on Binary Member Identifiers) is proposed to store frequent itemsets by using a prefix tree. Any node of this tree represents a frequent itemset and has an index to identify ”implicitly” the associated member Identifiers (MIDs). This index (Support of Binary member Identifiers) represents the member identifiers (MIDs) for the corresponding itemset. That is, suppose the data for a given problem are encoded as bit vectors of length n. Then any subset of the vectors can be represented by a boolean function over variables yielding 1 when the vector corresponding to the variable assignment is in the set. In our case, any member identifier can be encoded with $b_1,b_2,...b_2^m$ bits, where each bit b_k is in {0, 1} and m is the total number of members. The first level of TBBMI contains frequent 1-itemsets. For each node, we carry out the intersection of its TBBMI with the TBBMI of brothers nodes of same level to generate the second level with new frequent itemsets. We can say that each level k of TBBMI represents frequent k-itemsets. For instance, the transaction database as table I could be changed into a prefix tree as Figure 5.

TABLE I.

TRANSACTION DATABASE

Member ID	Transactions
M_1	A1,B2,C3,D1,E1
M_2	A2,B2,C3,D2,E2
M_3	A1,B2,C3,D3,E3
M_4	A3,B1,C1,D2,E1
M_5	A1,B3,C3,D2,E2

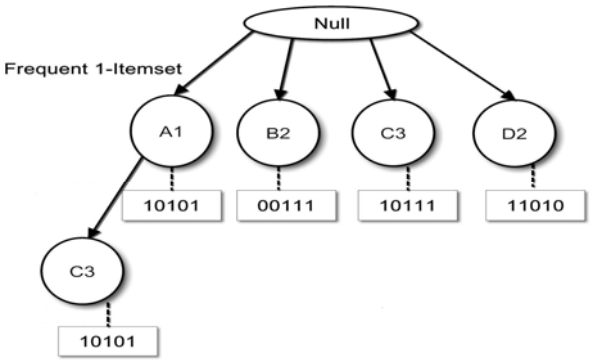

Figure 5. Tree Based on Binary Member Identifiers

Where TBBMI of A1 is ‘10101’, which means that member M_1, member M_3 and member M_5 all have the attribute ‘A1’. Provided that the minimum support of the example is 3, then the frequent 2-itemsets could be generated by counting the number ‘1’ in TBBMI of each node. Notice that Boolean function from the TBBMI of A1 and the TBBMI of C3: 10101 AND 10111 = 10101. The result ‘10101’ yields three ‘1’, thus the frequent 2-itemset A1C3 is obtained.

The process is repeated until the whole tree is constructed and finally the frequent itemsets could be

acquired from the tree, which means all the biclusters could be obtained too.

3 The pseudo-code of biclustering based on the improved Apriori

The algorithm of the improved Apriori is shown as following:

1) Scan the database. Find frequent 1-itemsets, encode the member ids with binary code， add them to the prefix-tree;

2) Join the frequent 1-itemsets into frequent 2-itemsets and add them to the prefix tree;

3) Obtain frequent k-itemsets: Find the leaf nodes with deepest depth in the prefix tree, Do (AND) operation on TBBMIs of its sibling leaf nodes, add it to the prefix-tree if the result meets the minimum support. Repeat the process until the new frequent itemsets could not be found.

4) Decode the binary code of the rows. The biclusters are the ouput of the prefix-tree.

The pseudo-code is stated as following:

- *Input：*

  1) Transaction database D (id,itemsets)， where id is member ID, itemsets is the subset of corresponding value of each attribute;

  2) Minimum support min_sup

- *Output：All submatrixes corresponding to the frequent itemsets*

```
/* reading and extracting the 1-itemsets*/
For each (Transaction T∈ D)do
  —For each (X ∈ T) do
  ——Add_items(X. Li);
    End For;
End For;
/* remove infrequent itemsets*/
Remove_item(Li,miniup);
/* Prefix-tree construction*/
frequent_itemsets=Create_Prefix_Tree(Li,minisup);
Return frequent_itemsets;
End
```

III EXPERIMENTAL RESULTS

The proposed method is validated through a dataset from the website of a domestic airline company which has 62988 members and 63 attributes and implemented on the computer with Win7 operating and Matlab7.0 platform whose RAM is 4G and CPU is AMD Athlon Ⅱ X4 645. To compare with our system, we implemented a traditional method (K-means clustering based on Recency-Frequency-Monetary model) [28] on the dataset.

A. Results of Traditional K-means Clustering

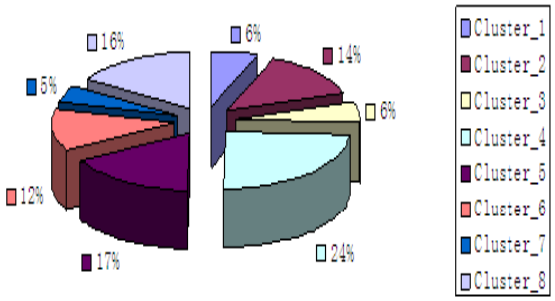

Figure 6. The clusters of the traditional K-means Algorithm

The customers are divided into 8-groups by the K-means clustering based on RFM model, and Figure 6 shows the fraction of each cluster.

The detailed information after K-means clustering is shown as Table II. Where each row represents the attributes related to the RFM model and the columns are the clusters. The entries of the table are the average values of the cluster on the corresponding attribute. For example, the element of the first row and the first column is 40.88, which means the average age of cluster_1 is 40.88. It is obvious that one could hardly acquire any more useful information through table II. Whereas the framework presented in the paper could do a good job.

B. Results of the Proposed Algorithm

After data preprocessing, there are 7 attributes left and the value of each attribute is quantized as 1, 2, 3 representing low, medium and high level respectively. The left attributes are shown as table Ⅲ.

TABLE II.

THE DETAILED INFORMATION OF 8 CLUSTERS

	Cluster_1	Cluster_2	Cluster_3	Cluster_4	Cluster_5	Cluster_6	Cluster_7	Cluster_8
Age	40.88	38.51	36.82	39.85	38.72	37.59	39.9	39.24
Age_Of_Card	5.48	3.55	3.33	4.11	3.71	2.82	2.86	3.47
Average_Discount	0.76	0.65	0.61	0.69	0.65	0.61	0.45	0.63
Price_Per_Kilo	0.76	0.66	0.62	0.69	0.65	0.66	0.73	0.64
Last_Day_Till_New	11.22	219.57	226.3	35.37	160.07	77	144.16	81.8
Flightratio_holiday	0.34	0.01	0.95	0.36	0.31	0.85	0.37	0.07
Flightratio_weekend	0.15	0.07	0.04	0.19	0.34	0.08	0.22	0.14
Flightratio_workday	0.5	0.92	0.01	0.45	0.35	0.06	0.42	0.78
Group_Ratio	0.21	0.44	0.43	0.32	0.4	0.42	0.22	0.4
…								
GroupFlight_Count	35.7	1.81	1.57	11.03	3.55	3.11	3.51	4.76
Agent_booking_ratio	0.921	0.904	0.772	0.934	0.898	0.924	0.867	0.937
B2C_booking_ratio	0.007	0.067	0.199	0.012	0.069	0.02	0.009	0.011
Phone_booking_ratio	0.047	0.017	0.019	0.032	0.018	0.035	0.1	0.028

Card_paid_ratio	0.231	0.206	0.282	0.254	0.222	0.299	0.052	0.312
Other_paid_ratio	0.06	0.004	0.003	0.052	0.012	0.033	0.795	0.039
Cash_paid_ratio	0.709	0.79	0.715	0.694	0.767	0.668	0.153	0.649
…								

TABLE III.

THE EXPLANATION OF THE SELECTED ATTRIBUTES

Attribute name	Explanation
EXPENSE_SUM	The total ticket price (the first year's ticket price + the second year's ticket price)
ELITE_POINTS_SUM	Sum of the elite points
FLIGHT_COUNT	Count number of flight
SEG_KM_SUM	The total segments distance of the flight(Km)
WEIGHTED_SEG_KM	The weighted total distance of the flight(Σ discount × segment distance)
Points_Sum	Sum of the points
BASE_POINTS_SUM	Sum of the base points

TABLE IV.

BICLUSTERS OF THE FIRST CLASS

Customers' number	EXPFNSE_SUM	SEG_KM_SUM	WEIGHTED_SEG_KM	Points_Sum	FLIGHT_COUNT	BASE_POINTS_SUM	ELITE_POINTS_SUM
42502	1	1	1	1	1	1	1
43308		1	1	1	1	1	1
43476	1			1	1	1	1
42555	1	1	1	1	1		
43631	1	1		1		1	1
42507	1	1		1	1	1	

TABLE V.

BICLUSTERS OF THE SECOND CLASS

Customers' number	EXPENSE_SUM	SEG_KM_SUM	WEIGHTED_SEG_KM	Points_Sum	FLIGHT_COUNT	BASE_POINTS_SUM	ELITE_POINTS_SUM
1823	2	2	2	2	2	2	1
1580	2	2	2	1	2	2	1
2318	2	2	2	2		2	1
1817	2	2	2	1		2	1
…	…	…	…	…	…	…	…
1907	2	2	2	2	2		
963	1	2	2	1			1
1889	2	2	2	1	2		
4772	2	2	2				
2518	2	2	2	2			
2136	2	2	2	1			

TABLE VI.

BICLUSTERS OF THE THIRD CLASS

Customers' number	EXPENSE_SUM	SEG_KM_SUM	WEIGHTED_SEG_KM	Points_Sum	FLIGHT_COUNT	BASE_POINTS_SUM	ELITE_POINTS_SUM
138	2	3	3	2	3	3	
52	2	3	3	2	3	2	2
48	2	3	3	2	3	2	1
149	2	3	3	2	3	3	
233	2	3	3	2	3		
235	2	3	3	2			
37	2	3	2				

DBSCAN algorithm divide the customers into three class).The above Table Ⅳ, Table V, Table VI are biclusters of each class.

It could be seen from TableⅣ that the attributes of the crowd are in the low level where the backslash means the corresponding attribute is not taken into account. These people are less involved in the flight consumption. The

airline company needs to improve its services to increase airline sales amount.

The middle class customer group is shown in Table V, who can be regarded as customers with potential value. Some of them could step to customers with high value. Each row represents the behavior with different characteristics.

For instance, the Points_Sum of the customers in the second row (Table V) belongs to a low level while the other attribute such as consumption, flight number, and flight distance have reached a good level.

The accumulated points (Table V) are not so important among the people of the class. Customers with low points gave much contribution to the air company as customers in the last three rows do. Therefore they should not be treated differently.

The base points of the customers in the third class (Table VI) have reached the highest level as well as the flight times and the flight distance. But there is still room for improvement considering about the elite points, the level of flight consumption and the total points.

The consumption level of this class (Table VI) should be the highest, while the consumption level of 295 customers (among total 331 customers) focused on the value of 2 (medium level). So it is necessary to give more preferential policies to stimulate their desire for consumption.

As the high end users, most of the BASE_POINTS_SUM (points) (Table VI) are the medium level, thus the professional work for the basic point should be strengthened.

We also compared the performance of the biclustering on Apriori with the biclustering on improved Apriori algorithm. The results are shown as Figure7.

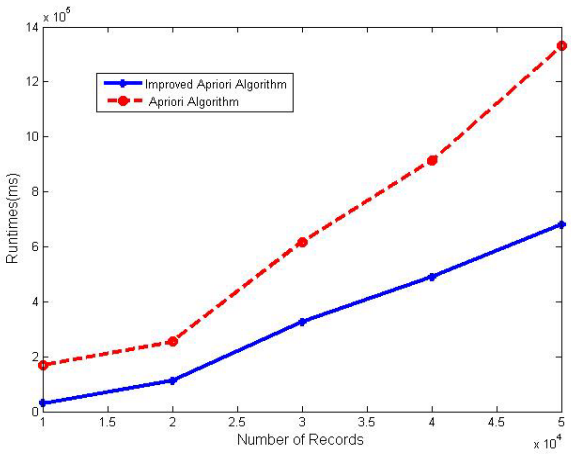

Figure 7. The comparison between improved Apriori and the original Apriori

It could be obviously discerned from Figure7 that the improved algorithm greatly reduces the number of database access and the searching cost as the data getting bigger, which greatly improved the efficiency.

IV CONCLUSIONS

In this paper, a novel approach for customer segmentation is proposed. Firstly, the chi-square statistical analysis is applied to choose set of attributes and K-means algorithm is employed to quantize the value of each attribute, then DBSCAN algorithm based on density is introduced to divide the customers into three groups (the first, the second and the third class), finally biclustering based on improved Apriori, which is used in the three classes to obtain more detailed information. Results of the experiment show that the biclustering could segment the customers more accurately and meticulously. A comparison was also made between the traditional K-means algorithm and the presented method, which proved that the latter could be more preferable than the former.

ACKNOWLEDGMENT

The authors thank gratefully for the colleagues who have been concerned with the work and have provided much more powerfully technical supports. The work is supported by Guangdong Science and Technology Department under Grant No.2009B090300336, No.2012B091100349; Guangdong Economy & Trade Committee under Grant No. GDEID2010IS034; Guangzhou Yuexiu District science and Technology Bureau under Grant No 2012-GX-004; National Natural Science Foundation of China (Grant No:71102146, No.3100958); Science and Technology Bureau of Guangzhou under Grant No. 2011J4300046; Research supported in part by the Guangzhou Research Infrastructure Development Fund (No. 2012224-12), the Guangdong Nature Science Fund (No. S2012030006242), the Guangzhou Zhujiang Future Fellow Fund (No. 2011J2200089), China MOE Doctoral Research Fund (No. 20134407120017), the MOE-China Mobile Research Fund (No. MCM20121051).

REFERENCES

[1] Alex Berson Ste.Building Data mining applications for CRM. Posts & Telecommunications Press, Beijing, 2001 (in Chinese).

[2] Jaesoo Kim, Sherrie Wei, Hein Ruys (2003). “Segmenting the Market of West Australian Senior Tourists Using an Artificial Neural Network”. Tourism Management, No. 24, pp. 25-34.

[3] Meng Xiaolian, Yu Yang (2007). “The Study on the Customer Identifying Model Based on Customer Value in Commercial Banks”. Value Engineering, No.7, pp.6-10 (in Chinese).

[4] M. McDonald. The role of marketing in creating customer value. In Marketing from an Engineering Perspective (Digest No. 1996/172), pages 1/1–111, Nov 1996.

[5] Guozheng Zhang. Customer segmentation based on survival character. In WiCom, Sept. 2007.

[6] Jing Wu and Zheng Lin. Research on customer segmentation model by clustering. In ICEC, New York, NY, USA, 2005.ACM.

[7] Girish Punj and David W. Stewart. Cluster analysis in marketing research: Review and suggestions for application. Journal of Marketing Research, 20(2):134–148, 1983.

[8] Vasilis Aggelis and Dimitris Christodoulakis. Customer clustering using rfm analysis. Ste-vens Point, Wisconsin, USA, 2005. WSEAS.
[9] W. Niyagas, A. Srivihok, and S. l Kitisin. Clustering e-banking customer using data mining and marketing segmen-tation. ECTI, 2(1), May 2006 .
[10] He Ling,Wu Lingda. " Survey of Clustering Algorithms in Data Mining, " Application Re-search of Computers, 2007(1)55-57.
[11] P. N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, Pearson Addison Wesley, Boston, 2006.
[12] J. Han, J. Pei, and Y.Yin, "Mining frequent patterns without candidate generation", In: Proceedings of the 2000 ACM-SIGMOD International Conference on Management of Data, Dallas, 2000, pp.1-12.
[13] CHENG Yizong, GEORGE M Chureh. Biclustering of expression data[C], Proceeding of the 8th International conference on Intelligent SystemS for Molecular Biology (ISMB00). MenloPark, California: TheAAAIPress, 2000: 93-103.
[14] Rakesh Agrawal and Ramakrishnan Srikant, Fast algorithms for mining association Rules in Large Databases, Proceedings of the 20th International Conference on Very Large Data Bases(VLDB'94), Santiago de Chile, pp.487–499, 1994.
[15] Aaron Ceglar and John F. Roddick, Association mining, ACM Computing Surveys, New York, Vol.38, No.2, 2006.
[16] Jiawei Han, Hong Cheng, Dong Xin and Xifeng Yan, Frequent pattern mining: current status and future directions, Data Mining Knowledge Discovery, Vol.15, No.1, pp.55-86, 2007.
[17] Gaurav Pandey, Gowtham Atluri, Michael Steinbach, Chad L. Myers and Vipin Kumar, An Association Analysis Approach to Biclustering, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, pp.677-686, 2009.
[18] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, Burlington, Massachusetts, 2011.
[19] Yong Li, Applications of Chi-square test and contingency table analysis in customer satisfaction and empirical analyses, 2009 International Conference on Innovation Management.
[20] [20] Guillem Lefait, Tahar Kechadi,Customer Segmentation Architecture Based on Clustering Techniques, 2010 Fourth International Conference on Digital Society.
[21] Yaqi Zhang, Chunhua Yin, Gang Zhao,FP-Tree Algorithm Applied in Logistics Warehousing Mining, 2011 International Conference on Electronic & Mechanical Engineering and Information Technology.
[22] Wassim AYADI, Khedija AROUR, A Novel Parallel Boolean Approach for Discovering Frequent Itemsets, Seventh IEEE International Conference on Data Mining – Workshops,2007,11
[23] Liu Wei, Chen Ling,A Fast Mining and Updating Algorithm for Frequent Patterns on Biological Single Sequence, 010 International Forum on Information Technology and Applications.2010
[24] Yo-Ping Huang and Li-Jen Kao , Frode-Eika Sandnes, A Prefix Tree-based Model for Mining Association Rules from Quantitative Temporal Data , Systems, Man and Cybernetics, 2005 IEEE International Conference on,2005,10
[25] W. Ayadi and Kh. Arour. A binary decision diagram to discover low threshold support frequent itemsets. In International Workshop on Advances in Conceptual Knowledge Engineering (ACKE'07), Regensburg,Germany, September 2007.
[26] Randal E. Bryant. Graph-based algorithms for boolean function manipulation. In IEEE, editor, Proceedings IEEE transactions computers,pages 677–691, August 1986.
[27] R.E. Bryant. Binary decision diagrams and beyond : Enabling technologies for formal verification. In proceedings of the International Conference on Computer-Aided Design, pages 236–243, November,1995.
[28] http://en.wikipedia.org/wiki/RFM_(customer_value)

Xiaohui Hu is currently an associate professor at School of Physics and Telecommunication Engineering, South China Normal University in China. Her research interests include pattern recognition, data mining, knowledge-based systems, etc. She received Bachelor's degree in Beijing Institute of technology, and Master's degree in Hunan University.

Haolan Zhang is currently an Associate Professor at NIT, Zhejiang University in China, and prior to that, he was a Research Fellow and academic staff at RMIT University, Australia. He has published over thirty papers in refereed journals and conferences. His research interests include Intelligent Information Systems, Multi-agent Systems, Health Informatics, Knowledge-based Systems, Data Mining, etc. He serves as the editorial board member and reviewer of several journals including: IEEE Transaction on SCM, The Computer Journal, and Journal of AAMAS. He serves as publicity chair, publication chair and PC members of several conferences. He is the Editor-in-Chief of Advances in Information Sciences, Guest editor of Journal of Internet Technology, etc.

Xiaosheng Wu now is an undergraduate in School of Physics and Telecommunication Engineering,South China Normal University. His research interests include data mining, cloud computing, machine learning pattern recognition, etc.