Clustering e-Banking Customer using Data Mining and Marketing Segmentation

Waminee Niyagas[1], Anongnart Srivihok[2], Sukumal Kitisin[3]

Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand
[1] waminee@scb.co.th, [2] anongnart.s@ku.ac.th, [3] fscismi@ku.ac.th

ABSTRACT

In Thailand e-banking has been offered by various financial institutes including Thai commercial banks and government banks. However, e-banking in Thailand is not widely used and accepted as in other countries. Accordingly, the study of e-banking is scantly due to the limitation of data confidentiality. This study uses data mining techniques to analyse historical data of e-banking usages from a commercial bank in Thailand. These techniques including SOMS, K-Mean algorithm and marketing techniques-RFM analysis are used to segment customers into groups according to their personal profiles and e-banking usages. Then Apriori algorithm is applied to detect the relationships within features of e-banking services. Typically, results of this study are presented and can be used to generate new service packages which are customised to each segment of e-banking users.

Keywords: RFM analysis, K-Means, SOMS, Apriori algorithm, e-banking, Market segmentation

1. INTRODUCTION

In these days, business needs to satisfy customer's demand to stay competitive. Consumer or customer is one who determines the direction of the market by buying products or services that satisfy him/her the most. If products or services which customer currently buys, no longer meets his/her increasing needs, he/she will choose products or services of other producers. If business is unable to understand behaviours of its customers, it soon will lose revenue and customers. Providing products/services that customers need without customers' request is one way to serve customers' needs. To be able to do so, business has to understand behaviours of its customers. Information of behaviours of customers can come from customer data that business has collected. When business understands behaviours and attributes of customers, it will be able to develop products or services that satisfy customers' demands.

Banking is one of highly competitive business. Banks are attempting to create channels or products that will help distinguish themselves from competitors. To be different from competitors and be able to satisfy customer's needs, e-banking or Internet banking is one channel that gives customers convenience and reduces customers' costs of travelling. Internet banking has gained popularity and has been used widely in many countries. But in Thailand, Internet banking has not been so successful [1]. In addition to Thais' resistance to change, and concern over lack of security of the Internet, limitations in Internet banking services of Thai banks such as unavailability of critical services. Thai customers can do only some functions of banking activities on the Internet. As well, services provided are not corresponding to customers' needs. A study of behaviours of Internet banking customers such as personal information or usages is one mean to help banks keep their existing customers and gain new customer base effectively [2]. Also to help banks identify what they lack and what should be added to serve customers' needs better.

Therefore, this study will demonstrate methods of studying customers' behaviours by applying data mining technique and RFM Analysis [3] in order to find valuable customers and their usage behaviours in order to add new products or services to meet customers' needs. Moreover we can use information derived from data mining to apply to marketing for e-banking.

2. DATA MINING

There are three main algorithms applied in this study.

2.1 K-means Algorithm

K-means algorithm [4-5] is the simplest clustering algorithm and widely used. K-means requires an input which is a predefined number of clusters. This input is named k. The steps of the K-means algorithm are given below.

1. Select randomly k points to be seeds for the centroids of k clusters.
2. Assign each point to the centroids closest to the point.
3. After all points have been assigned, recalculate new centroids of each cluster.
4. Repeat step 2 and step 3 until the centroids no longer move.

2.2 Kohonen's Self-Organizing Map (SOM)

SOM [6] is one of the most popular and powerful neural networks in the unsupervised learning domains. The basic idea is to :

1. Represent high dimension data in a low-dimensional form without loosing any of the 'essence' of data.
2. Organizes data on the basis of similarity by putting entities geometrically close to each other.

Summary Steps is below [7]:

1. Initialize the weight vectors
2. Decide on appropriate *g* and *r* values
3. For each input node
   - Find the shortest distance to any output node.
   - Adjust selected node's weight according to current stage of *g*.
   - Adjust neighboring node's weight according to current stage of *g*
   - Go to next unvisited input node. If there are no unvisited input nodes left then go back to the very first on and go to Step 4
4. When a previously visited node has been reached, incrementally decrease *g* and *r* and repeat Step 3.
5. Keep doing Steps 3 and 4 for sufficient number of iterations.

2.3 Apriori Algorithm

Association rules are among the most popular representations for local patterns in data mining. Apriori algorithm [8] is one of the earliest algorithms used for finding association rules. The algorithm is an influential algorithm for mining frequent item sets according to Boolean association rules. The pseudo code for Apriori algorithm is given below:

```
Ck : candidate itemset of size k
Lk : frequent itemset of size k
L1 = {frequent item};
For (k=1; Lk != null; k++) do begin
    Ck+1 = candidates generated from Lk;
    For each transaction t in database do
      Increment the count of all candidates in
      Ck+1 that are contained in t
    Lk+1 = candidates in Ck+1 with min_support
    End
Return Lk;
```

3. MARKETING SEGMENTATION

RFM analysis [8-9] is a three-dimensional way of classifying, or ranking, customers to determine the top 20%, or best, customers. It is based on the 80/20 principle that 20% of customers bring in 80% of revenues. RFM Analysis is a marketing technique that uses three features include Recency, Frequency and Monetary Value of customers to predict whether or not they are likely to buy again. Essentially RFM analysis suggests that the customer with high RFM score should normally conduct more transactions and result in higher profit for the bank [8-10]

The following features are calculated for this Specific time period.

Recency(R): R is the date of the user's last transaction.

Frequency (F): F is defined number of financial transactions that user conducted within specific period.

Monetary (M): M is the total value of financial transactions that user made within the above stated period.

RFM Score is calculated using the formula:

$$RFM = R + F + M \quad (1)$$

4. EVALUATION OF CLUSTERING

Three methods used for evaluating the efficiency of data segmentation are as follows.

4.1 Standard Deviation (SD)

The standard deviation is the most commonly used for measuring the variation of values in the defined dataset. The lower SD value means the better clustering.

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \text{ where } \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (2)$$

4.2 Root Mean Square Standard Deviation (RMSSTD)

The RMSSTD [11-12] is the variance of the clusters; RMSSTD measures the homogeneity of the clusters to identify homogenous groups, the lower RMSSTD value means the better clustering.

$$RMSSTD = \sqrt{\frac{\sum_{i=1..n_c, j=1..d}\sum_{k=1}^{n_{ij}}(x_k - \bar{x}_j)^2}{\sum_{i=1..n_c, j=1..d}(n_{ij}-1)}} \quad (3)$$

Where n_c is number of cluster, d is number of dimension

$\bar{x}_j$ is expected value in the j^{th} dimension.

n_{ij} is number of element in i^{th} cluster j^{th} dimension.

4.3 R Squared (RS)

RS [11-12] is used to measure the dissimilarity of clusters. Formally it measures the degree of homogeneity degree between groups. The values of RS range for 0 to 1 where 0 means there is no difference among the clusters and 1 indicates that there are significant difference among the clusters.

$$RS = \frac{SS_t - SS_w}{SS_t}, \text{ where} \quad (4)$$

$$SS_t = \sum_{j=1}^{d}\sum_{k=1}^{n_j}(x_k - \bar{x}_j)^2, SS_w = \sum_{i=1..n_c, j=1..d}\sum_{k=1}^{n_{ij}}(x_k - \bar{x}_j)^2$$

5. EXPERIMENT

5.1 Data set

The dataset of this study is Internet Banking customer data from one commercial bank in Thailand between January 1st and December 11th of the year 2005. There are 458,000 transactions of 2,096 active customers. The term «active e-banking user» describes the user who conducted at least one financial transaction during this period.

Five factors used in data segmentation included: (1) Date, (2) Time, (3) Status of Transaction, (4) Type of Transaction, and (5) RFM Score.

5.2 Framework

The framework of data mining in this study is depicted in Figure 1.

1. Select data based on customer profiles, transaction details, event log, and bill payment service provider. Then, remove data redundancy and organize data.
2. Calculate data for RFM, which rates each customer's importance, based on latest transactions, business volume and frequency of use
3. Clustering data by combining algorithm between SOM and K-Means from Clustering Technique
   Step 1: Use Kohonen's self-organizing maps neural network (SOM) Grouping data into 2-10 groups and find Standard Deviation of the each grouping to point out which group has the lowest SD. The lowest

one means good grouping, and least data dispersion. For selecting the best number of cluster, we use RMSSTD and RS). Then, the output from this step is used as input for the next Step.

Step 2: K-Means Algorithm is used for grouping data. The number of group or K value is derived from step 1. In this step customer data are segmented based on customer transactions and their behaviors.

4. Use Associate Rule technique to find the relationships between features of e-banking transactions in order to understand customer behaviours.

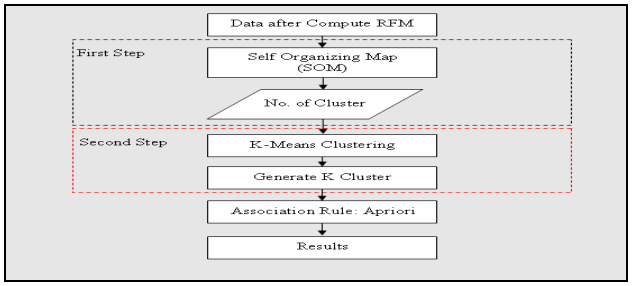

Fig.1: *System Framework for the study*

5.3 Results

First Step: By using self-organizing maps (SOMs) to cluster, we found 8 clusters is the best number of clustering. Fig.2 depicts the standard deviation of clustering ranging from 2 clusters to 10 clusters. Further, results from clustering data by SOMs indicate that SD decreases when the number of clusters increases.

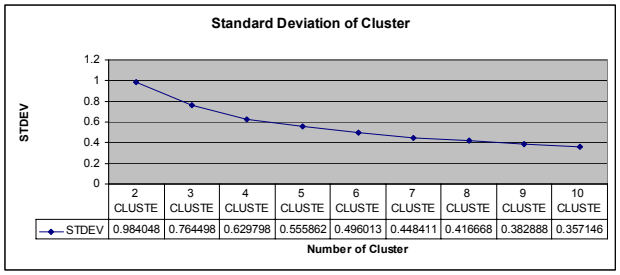

Fig.2: *Standard Deviation of different numbers of clusters*

In order to confirm the results from SD value, 'RMSSTD' and 'RS' techniques are used to measure similarities and differences among different groups. The result suggested that the optimal number of groups is 8 as shown in both Fig. 3 and 4.

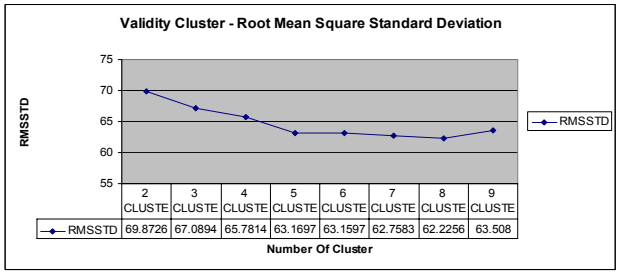

Fig.3: *Validity Measurement – RMSSTD*

Figure 3 Shows that the lower the RMSSTD is, the more difference the groups are. The optimal number of groups is 8 which results in the lowest RMSSTD of 62.2256.

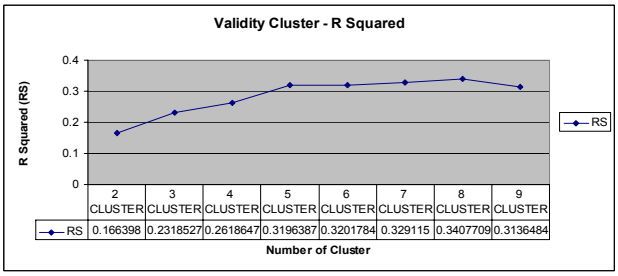

Fig. 4: *Validity Measurement – RS*

Figure 4 shows that the higher the RS is, the more difference the groups are. Figure 4 shows that the optimal number of groups is 8 of which the highest RS of 0.34077. Result is similar to the measurement by RMSSD.

Thus, in the first phase, the optimal number of groups (k) generated by SOM is 8. This number can then be used as a parameter for second phase-segmentation. In this phase, K-Means algorithm is applied for e-banking data segmentation.

Table 1: *Result of clustering by K-means*

Factor		Cluster 1	2	3	4	5	6	7	8	Total
		52	574	192	19	129	26	5	1099	2096
Date	1 - 15 of Month	8856	13838	3960	7465	10352	3916	4483	6658	59528
	16 - 30 of Month	10122	15981	4725	7769	12561	4178	4575	8473	68404
Time	0.00 - 5.59	96	70	23	0	72	3	0	45	309
	6.00 - 11.59	8069	13353	3427	3881	9029	3530	234	5770	47293
	12.00 - 17.59	9266	14088	4470	9982	11982	4255	8645	8020	70708
	18.00 - 23.59	1547	2308	765	1371	1850	306	179	1296	9622
Type	RealTime	17520	23703	6747	12751	19751	3736	7332	10704	102244
	Schedule	1458	6116	1938	2483	3182	4358	1726	4427	25688
Transaction	Balance	12576	57159	47317	4594	22147	5184	1255	26017	176249
	Report	5064	23891	26618	1389	8340	2056	435	10273	78066
	Transfer	18682	17095	6762	14799	21878	817	9057	7389	96479
	Payment	296	12724	1923	435	1055	7277	1	7742	31453
RFM		13	11	11	14	13	13	15	7	

There are eight clusters in this data set.

Cluster 1 has a few members (2.48%). Customers use the system often throughout the month especially around 3rd and 4th week. Day time usage (6-17.59 hours) includes account inquiry and balance transfer. *Highlight:* Customers use the system regularly for financial transactions, high volume and high profit.

Cluster 2 has a lot of members (27.39%). Customers use the system often but complete a few transactions. Day time usage includes account inquiry and reports. *Highlight:* Use the system regularly for financial transactions. Average volume and average profit.

Cluster 3 has average members (9.16%). Customers use the system moderately. Day time usage includes account inquiry and reports. *Highlight:* Use the system moderately for financial transactions, high volume and high profit.

Cluster 4 has very few members (0.91%). Customers use the system regularly to complete a lot of transactions between 12.00 - 17.59pm. Transactions are normally set in advance, most are balance transfer. *Highlight:* Use

the system regularly for financial transactions, high volume and high profit.

Cluster 5 has average members (6.15%). Customers use the system frequently between 12.00 – 17.59pm. Transactions are normally set in advance. Most transactions are account inquiry and balance transfer. *Highlight:* Use the system frequently for financial transactions, high volume and high profit.

Cluster 6 has a few members (1.24%). Customers use the system occasionally between 12.00 – 17.59pm. Most transactions are payment and account inquiry. *Highlight:* Use the system occasionally for financial transactions. High volume and average profit.

Cluster 7 has the fewest members (0.24%). Customers use the system infrequently but consistently complete a lot of transactions between 12.00 – 17.59pm. Most transaction is balance transfer. *Highlight:* Important customers use the system frequently for financial transactions, very high volume and very high profit.

Cluster 8 has the maximum members (52.43%). Customer seldom use the system complete a few transactions. Most usage occurs between 12.00 – 17.59pm. Most transactions are account inquiry and reports. *Highlight:* Seldom use the system for financial transactions, low volume and low profit.

Applying Apriori Algorithm

The relationships between different payment types of the customers of each cluster were determined by using Apriori algorithm. The results are follows. Clustering 1: Customers using 3rd party transfer services are also use Account transfer services.

Clustering 2: Customers using Direct Credit services are also use 3rd party transfer and Account transfer services.

Clustering 3: Customers using Media Clearing and 3rd party transfer services are always use Account transfer services. Customers using BAHTNET services are also use Media Clearing services.

Clustering 4: Customers using Payroll transfer services are also use Account transfer services.

Clustering 5: Customers using Payroll transfer services are also use 3rd party transfer and Account transfer services.

Clustering 6, 7, 8: Customer have similar behaviours they use 3rd Party transfer then use own account Transfer. Otherwise, the customer using Payroll then use own account transfer.

5.4 Discussion

The results show that daily e-banking concurrent access was increased during 12.00 – 17.59 pm., impact the system to serve all accessing with slowly response. For long terms improvement, the bank should consider to renovate the e-banking infrastructure: Network bandwidth, Server and Database capacity through monitoring tools. For short terms improvement, the bank should consider in e-banking customer behaviors to adapt their access behaviors by extended time access, promoted schedule for transactions, and promoted new packages to increase more revenues.

6. CONCLUSION

This research focuses on clustering e-banking customer to analyze customer characteristics and behaviors with appropriated criteria: access time, transaction access and RFM Analysis. The benefits are valuable for the bank to improve services. The research shows distinct clustering results as follows: The relationship of financial transaction as transfer was significant both group 8 (Largest Cluster - 52.43%) and group 7 (lowest cluster -0.24%), if customers use 3rd Party transfer then they always use their own account transfer. As well, if customers use payroll then they always use own account transfer. However, RFM Analysis is considered to show most of customer in group 8 is inactive and imply to make less valuable for bank while group 7 make more valuable for bank even. These results might be benefit for banking marketing team to launch a suitable promotion for appropriate clustering.

7. REFERENCES

[1] M. Ongkasuwan and W. Tantichattanon, "A Comparative Study of Internet Banking in Thailand." *The First National Conference on Electronic Business* , 2002.

[2] S. Wiwattanacharoenchai and A. Srivihok., "Understanding online Banking in Thailand: Cluster Analysis of customer usage behaviour". *Asia-Australasian Regional Conference*. 22-24 June 2003.

[3] DataPlus Millenium, "*Data-Driven Analysis Tools and Techniques*", White Paper, 2001.

[4] P. Bradley and U. Fayyad., "Refining Initial Points for K-Means Clustering". Proc. *15th International Conf. on Machine Learning*, 1998.

[5] U. Fayyad, S. G. Piatetsky and P. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework". *The AAAI press*. 156p., 1996.

[6] Kohonen, T. Self-Organizing Maps. Berlin-Heidelberg, Springer.

[7] http://www.ucl.ac.uk/oncology/MicroCore/HTML_resource/SOM_Ini.htm

[8] V. Aggelis, "RFM analysis with Data Mining", Scientific Yearbook, Technological Education Institute of Piraeus, 2004.

[9] V. Aggelis and D. Christodoulakis, "RFM analysis for decision support in e-banking area", *WSEAS Transactions on Computers Journal.* ISSN 1109-2750, 2005.

[10] J. Galindo., "Credit Risk Assessment using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications", Computational Economics Journal, 1997.

[11] A. K. Jain, M. N. Murty and P. J. Flynn, *"Data clustering: a review"*, ACM Computing Surveys, Vol. 31, No. 3, pp. 264-323, 1999.

[12] M. Halkidi, Y. Batistakis and M. Vazirgiannis, *"Cluster validity methods: part II"*, SIGMOD Rec., Vol. 31, No. 3, pp 19-27, 2002.