한국지능정보시스템학회논문지 제13권 제4호 2007년 12월(pp.17~34)

Rethinking of Self-Organizing Maps for Market Segmentation in Customer Relationship Management*

Jounghae Bang
College of Business Administration
Kookmin University,
(bangjh@kookmin.ac.kr)

Lutz Hamel
Department of Computer Science and Statistics,
University of Rhode Island, USA
(hamel@cs.uri.edu)

Brian Ioerger
Department of Computer Science and Statistics,
University of Rhode Island, USA
(ioerger@worldnet.att.net)

Organizations have realized the importance of CRM. To obtain the maximum possible lifetime value from a customer base, it is critical that customer data is analyzed to understand patterns of customer response. As customer databases assume gigantic proportions due to Internet and e-commerce activity, data-mining-based market segmentation becomes crucial for understanding customers. Here we raise a question and some issues of using single SOM approach for clustering while proposing multiple self-organizing maps approach. This methodology exploits additional themes on the attributes that characterize customers in a typical CRM system. Since this additional theme is usually ignored by traditional market segmentation techniques we here suggest careful application of SOM for market segmentation.

Keywords : CRM, Market Segmentation, Self-Organizing Maps, Unsupervised Learning, Clustering, Online-Marketing

논문접수일 : 2007년 03월 게재확정일 : 2007년 12월 교신저자 : 방정혜

1. 서론

Organizations have realized the importance of Customer Relationship Management (CRM) (Kotler, 1997; Reichheld and Sasser, 1990). To seek the best possible lifetime value from customers, it is important to respond effectively to customer requirements (Luxton, 2002). For this, it is paramount that customer data is analyzed to understand customer response patterns. Business-to-con-

* This work was supported by the faculty research program (2007) of Kookmin University in Korea

sumer (b2c) multi-channel marketers, with e-commerce an increasingly vital channel, collect data on customers' behaviors, preferences, and their perceived value to the organization. Such multi-channel marketers, however, face the challenge of leveraging this data for marketing campaigns. In this era of the Internet and e-commerce, customers expect better targeted and personalized marketing communications with the right frequency. Customers are all too eager to shift loyalty if dissatisfied (Luxton, 2002).

Since most organizations have thousands if not millions of customers, the question becomes: "Whom to target with what?" Despite the increasing investment in CRM application systems (IDC, 2001), organizations continue to be unsuccessful in understanding their customers due to the failure of turning massive data warehouses into actionable information (Luxton, 2002). In our view this is partly a question of linking market segmentation to data mining methods. Past studies have investigated how data mining can be used to better understand customers from CRM perspectives. Data mining methods can be employed to find segments of customers who want a relationship with an organization (Bang et al. 2004). Self-organizing maps (SOM) (Kohonen, 2001) has been known as one of the most common data mining tools for clustering (Shaw et al., 2002) and used for research on market segmentation which was performed over one theme such as customers' shopping patterns and decision making styles. However, there is possibility that additional structure exists within the collection of attributes that characterize customers in a CRM data warehouse due to rich data collection. For example, in a typical data warehouse the shopping behavior of a customer is not expressed as a single attribute but as a group of attributes that describe the activity of the customer at various touch points (including web based contacts), the kind of products pursued in various channel (including the Internet), and the kind of customer service calls fielded amongst many others. This possible additional structure may not be revealed by a single SOM approach. This study examines the single SOM approach to see if the multi-group structure can be revealed and introduces possible problems and issues while reviewing a multiple SOM approach.

The remainder of the paper is structured as follows. We first discuss market segmentation in general and define some of its most important parameters and characteristics. Overviews of previous literature and of our multiple SOM methodology followed by a detailed definition of our segmentation methodology are given. Finally we work through an example using our methodology employing a synthetically constructed customer database. We contrast our results with the results obtained by the single SOM approach of only constructing a single self-organizing map. We conclude the paper with some final remarks and observations.

2. Background

2.1 Market Segmentation and CRM

One of the major advantages of CRM is that

by using CRM organizations can leverage the continuous stream of customer data collected through various touch points, facilitating individual level marketing decisions (Libai et al., 2002). With e-commerce, such data streams have become raging torrents, as every "mouseover" and click in a web-based customer interface can potentially be logged. According to Ryals (2003), CRM has three objectives: increase in customer value and satisfaction, identification of the most valuable customers or segments, and reduction of the risk of valuable customers defecting to competitors and target them successfully with retention strategies. The second of these three CRM objectives is of course the familiar market segmentation approach, a mainstay of most marketing strategies.

Segmentation has long been discussed in the marketing literature because of the heterogeneity among consumer preference (Kardes, 1999). Marketers can offer different marketing programs such as communications, distribution, and products to different groups of consumers if marketers can find the groups of consumers which show difference between groups, but similarity within each group (Schewe and Meredith, 2004). Many studies have been conducted on market segmentation, using several different bases for segmentation. For example, Walsh et al. (2001) used the consumers' decision-making styles, Papatla and Bhatnagar (2002) used consumers' shopping styles, and Mitchell (2003) used the consumers' attitudes towards unsolicited direct mail and telesales as bases for segmentation. Those studies used one theme for segmentation.Not only just one theme to segment customers, but also can combinations of segmentation themes be used if the right data exists (Berry and Britney, 1996). For example, in a banking setting, one theme could be channel preference, which classifies customers into segments based on their relative use of the bank's services and sales channels-including of course the growing use of Internet banking. Another theme could be transaction frequency based on the total volume of transactions over a fixed period of time. Because of the complexity of database and availability of data mining tools, several combinations of segmentation themes, that are additional structure in the database, can be employed to explore latent patterns and structures. In this study, combination of value and behavioral themes are studied instead of using just one segmentation theme.

2.2 Data Mining Tools for Market Segmentation

As a tool to analyze CRM-related customer data, data mining has received substantial attention especially as huge databases are being generated via the Internet (Fayyad et al., 1996; Mackinnon and Glick, 1999). Data mining can be seen as one step in the process of knowledge discovery in databases (KDD); the oft-iterative process of data selection, sampling, preprocessing, cleaning, transformation, dimension reduction, analysis, visualization, and evaluation (Mackinnon, and Glick, 1999). Many studies have been conducted on data mining techniques and marketing decision-making. For example, Koh and Chan (2002) showed all the

possible ways of using data mining techniques for CRM in the banking industry and Kim, Kim and Lee (2001) proposed a methodology to enhance the accuracy in predicting the tendency of customer purchase behavior by combining multiple classifiers based on genetic algorithms (Goldberg, 1989). Shaw et al. (2002) introduced three major areas of application of data mining for knowledge-based marketing: (1) customer profiling, (2) deviation analysis, and (3) trend analysis. Here we employ unsupervised learning in the form of self-organizing maps to extract knowledge from a customer database. The knowledge is in the form of segments or clusters of customers (Shaw et al., 2002).

2.2.1 Unsupervised Machine Learning : Self-Organizing Maps and Market segmentation

Self-organizing maps (Kohonen, 2001) were introduced by Kohonen in 1982 and can be viewed as tools to visualize structure in high-dimensional data (Witten and Frank, 1999). Self-organizing maps map n-dimensional input data to a two-dimensional plot (Kiang and Kumar, 2001). However self-organizing maps "preserve the underlying properties of the data" (Curry et al., 2003). Therefore, self-organizing maps have been studied as a tool to mine web log data and to provide a visual tool to assist user navigation in online surfing (Smith and Ng, 2003) and as an alternative to factor analysis (Kiang and Kumar, 2001). In their study, Kiang and Kumar (2001) suggest that "self-organizing maps can provide solutions superior to unrotated factor solutions in general and provide more accurate recovery of underlying cluster structures when the input data are skewed." Self-organizing maps (Kohonen, 2001) are considered members of the class of unsupervised machine learning techniques, since they do not require a predefined concept but will learn the structure of a target domain without supervision (Witten and Frank, 1999). Shaw et al. (2002) address self-organizing maps as one of the most common data mining tools for clustering. A number of extensions have been developed for self-organizing maps over the years to make them more applicable to market analysis.

Recent studies of self-organizing maps have been found in the area of market segmentation. Curry et al. (2003) examined consumer attitudes toward direct marketing. In this study, the authors used self-organizing maps, as an alternative to cluster analysis, to explore the most important factors in consumers' purchasing behavior. The self-organizing maps were compared with clustering methods. They noted that clusters are constructed to be mutually exclusive while self-organizing maps allow neighboring nodes to bear some relationship to each other, having been collectively adjusted during the training process. They also argued that self-organizing maps have an advantage over clustering in the depiction of the data. Self-organizing maps can examine both differences and similarities across segments while clustering only sees the differences. A self-organizing map analysis shows a number of people who are on the borderline and "shows clearly how the respondent profile changes when moving from one segment to another (Curry et al., 2003)." Most extensions of

self-organizing maps in market segmentation entail a multi-stage process where the self-organizing map is considered a "preprocessor" that estimates the number and rough characteristics of existing segments in the data. This is followed by a stage that looks at the segments in more detail (Krieger and Green, 1996; Kuo et al, 2002). Kuo, Chang, and Chien (2004) studied self-organizing maps and clustering methods for market segmentation. They proposed the two-stage method in which self-organizing maps are used first to determine the number of clusters and then genetic-algorithm-based clustering method is used to find the final solution. Kuo, Ho, and Hu (2002) proposed a combination of the self-organizing maps and K-means method. They compared three clustering methods: (1) conventional two-stage method using multivariate analysis, (2) the self-organizing maps, and (3) two-stage method with self-organizing maps and K-mean algorithm. They recommended the use self-organizing maps for identifying number of clusters and K-mean algorithm for further investigation of the segmentation. One of the most in-depth studies of various extensions is Mazanec (1999)'s, which looks at self-organizing maps in the context of market structure analysis.

As seen above, self-organizing maps have been employed for market segmentation. In most cases, self-organizing maps have been used to identify the number of segments in the initial process, followed by the use of other methods because "the autoclustering feature of self-organizing map is more effective and objective than the K means method, allowing clusters of arbitrary size (Kuo et al., 2004)." These studies have utilized the ability of self-organizing maps to identify the segments under one segmentation theme. Now that databases and data warehouses become larger and many different types of data become available, there are more possibilities to utilize the segmentation approach for data which containing additional structures. Market segmentations address hierarchically structured data or rich data which contains many different themes. Single SOM approaches, however, have been found to lack the ability to extract the hierarchical structure of the data (Pampalk, Widmer, and Alvin Chan 2003). Advanced SOM approaches, such as hierarchical feature maps (Miikkulainen, 1990) and growing hierarchical SOM (Rauber, Merkl, and Dittenbach, 2002), have been developed for those hierarchically structured data where some attributes are nested under some other attributes. Those approaches are viewed as multi-layered SOMs approach (Rauber, Merkl, and Dittenbach, 2002; Miikkulainen, 1990). After predefining the granularity of the individual SOMs and the overall depth of the structure are used (Rauber, Merkl, and Dittenbach, 2002; Miikkulainen, 1990), a top-layer SOM identifies the structure of first layer data, which would be future explored on SOMs in sub-layers. Growing hierarchical SOM approaches uses a flexible structure which automatically grows to fit the data by either adding new row and column or expanding units on the next hierarchical level. Therefore these advanced hierarchical SOM approaches use all attributes in every SOM layer, which can reveal unbalanced sub-layers of data since some categories identified

in the top layer SOM can be explored into more sub-categories while some other categories in top layer may not have any other sub-categories (Pampalk, Widmer, and Alvin Chan 2003).

On the other hand, with the rich data available, it is also important to perform the market segmentation with two or more themes such as attitudes toward a brand and preferred shopping channels (Berry and Britney, 1996). In order to address this type of market segmentation, here we attempt to use (1) a single SOM approach first to see if it can reveal the multi-conceptual dimensions of data, (2) to use a multiple SOM approach which repeats SOMs for each theme borrowing the concept of hierarchical SOM approaches, and (3) to raise problems and issues of using single SOM approach and clustering method.

3. Overview of Multiple SOM Approach

In this section, the single SOM approach is briefly reviewed and a multiple SOM approach will be discussed.

3.1 Dataset and single SOM approach

Regardless of the exact parameters chosen for market segmentation, the attributes necessary for the characterization of customers need to be retrieved from some sort of data store, typically a data warehouse. Once retrieved, such data can usually be viewed as a table where the customers make up the rows or records of this table and the attributes that describe the customers make up the columns of table. [Figure 1] shows an outline of such a table. In [Figure 1] the customers 1 through K are shown as rows and each customer has specific values in the attribute fields 1 through N.

[Figure 1] Typical Customer Description Table for Market Segmentation

Customer ID	Attribute 1	···	Attribute N
Customer 1			
···			
Customer K			

In typical market segmentation applications this table is interpreted as an N-dimensional Cartesian space where each attribute represents a dimension and customers are points in this space. The goal of market segmentation is to identify meaningful clusters of customers in this N-dimensional space. This is accomplished with clustering algorithms such as k-means, hierarchical clustering, or self-organizing maps amongst others (StatSoft, 2004).

3.2 Multiple SOM approach

Although this single SOM approach to market segmentation is quite successful in identifying target clusters of customers it is questionable if this approach would reveal the segments when some substantial additional structure exists on the attributes. Consider for the moment that perhaps one of the criteria for segmentation is the shopping behavior of customers. In a typical data warehouse

this behavior is not expressed as a single attribute but as a collection of attributes. If additional theme is included and therefore another collection of attributes are added in, the single SOM approach may not be able to reveal the structure because of potential confounding effect of two separate groupings of attributes. This additional grouping of the attributes has not been taken into account by the single SOM approach to market segmentation. In single approaches each attribute is treated as a dimension in its own right, the possibility that certain attributes belong together and should be treated as a unit cannot be represented. We claim that this lead to a less than optimal customer segmentation as our experiments show below.

The multiple SOM approach to market segmentation takes this additional structure into account. Instead of clustering on all attributes at the same time, we only cluster on attributes that represent a natural grouping at the same time. [Figure 2] shows our customer table from Figure 1 with a set of groupings, i.e. behavior grouping, value grouping and opt-in grouping, imposed on it. Each grouping is a collection of related customer attributes and defines a subspace of the N-dimensional space defined by the original table. If we now impose some combination of the groupings and select the order for clustering, for example: behavior → value → opt-in, then we can cluster our customers first based on their behavior using only the attributes given in the behavior grouping. Given the behavioral clusters we can then cluster the customers in each behavioral cluster according to only their value attributes. And finally, we segment the value clusters by their opt-in attributes.

A consequence of this approach is that we potentially face an explosion of possible segmentations to be constructed at each level. For example, given our three groupings and assuming that each grouping gives rise to two customer clusters for each input population we are faced with constructing seven segmentations. [Figure 3] shows this situation graphically.

[Figure 2] Customer Description Table with the Corresponding Attribute Groupings

	Behavior Grouping	Value Grouping	Opt-In Grouping
Customer ID	*Attribute 1… Attribute B*	*Attr B+1… Attr V*	*Attr V+1 … Attr N*
Customer 1			
…			
Customer K			

As our experiments show, respecting the inherent additional structure on the attributes during market segmentation results in more accurate customer clusters than the single SOM methodology.

4. Methodology

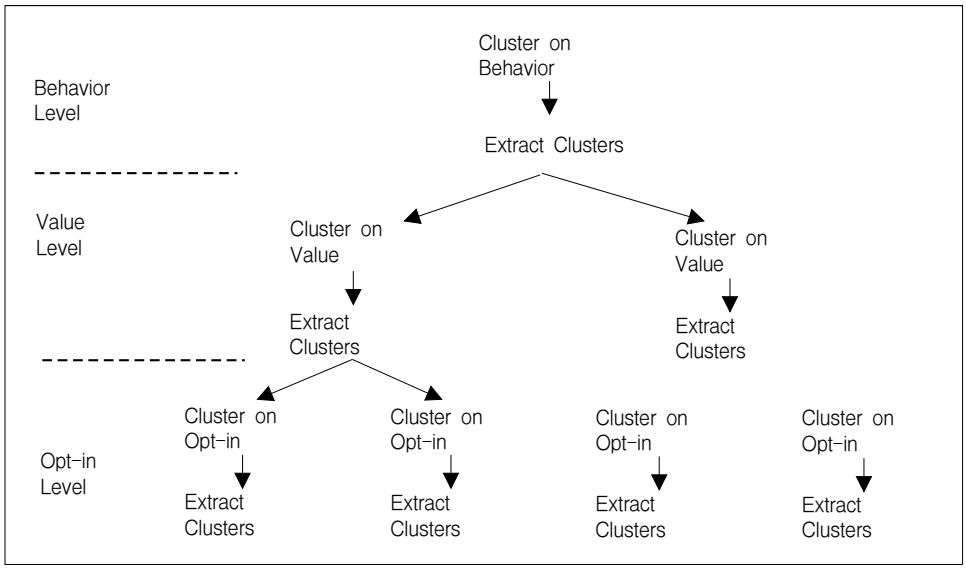
In order to present multiple SOM methodology in more concrete terms we summarize it here in quasi algorithmic format. Table 1 shows the steps necessary in the methodology and is a generalization of the situation shown in Figure 3. We sometimes refer to the groupings as *Conceptual Dimensions*, to highlight the fact that these are not

actual (original) attributes or dimensions of the data but represent additional synthesized, conceptual structure on the data.

The way the methodology is outlined in this table still suffers from the shortcomings mentioned in the previous section in that it forces the analyst to build all possible customer clusters. Campaign designers, however, typically have certain customer characteristics in mind and therefore not all possible clusters need to be consulted and the "search tree" in [Figure 3] can be pruned to something manageable.

In our experiment, we only used two themes (behavior and value) instead of using three themes (behavior, value and opt-in) to simplify the procedure. As well, in order to avoid this potential

[Figure 3] Visual Representation of Relationship between Levels in a Hierarchy and Number of Times Customer Clusters Need to be constructed



<Table 1> Methodology

1. Extract customer data table from data sources (data warehouse).
2. Group customer attributes according to their conceptual dimensions.
3. Make the order of the conceptual dimensions for clustering.
4. Starting at the first dimension; while there are other dimensions not processed do:
    a. Cluster each customer population with the first conceptual dimension.
    b. Extract the customer population for each resulting cluster; these become the input populations for the next iteration.
    c. Move to the next conceptual dimension.
5. Identify target cluster of customers.

combinatorial explosion in cluster construction, only the promising clusters at each dimension level are explored further during actual applications of this methodology.

4.1 Campaign Design in an Idealized Data Warehouse

Organizations typically collect data on many different attributes for each customer. The challenge in campaign design is to use these attributes in order to select the most promising sub-population of customers to target in the data warehouse. In order to study, compare, and contrast multiple SOM approach with the single SOM approach, we built a data generator that allows us to populate an idealized customer database which consists of non-overlapping populations. The discrete populations can be regarded as a simplest form of database for segmentation and lead to straightforward examination of the segmentation result of single SOM approach and the proposed approach.

Customer ID	Behavior Attributes	Value Attributes
...	...	...

[Figure 4] Our Simple Data Schema Using Conceptual Dimensions for Behavior and Value.

The data generator stores attributes on customers using the simple schema shown in Figure 4. The behavior attributes and the value attributes consist of multiple attribute fields describing be-

<Table 2> Typical Attributes Generated by the Data Generator for a Simulated Multi-Channel Bank

ATM	Online	Call Center	Walk-in
Interaction Mode *Banking Transaction: *Withdraw* *Deposit* *Inquiry* *Transfer* *Maintain*	Interaction Mode *Banking transaction: *Transfer* *Inquiry* *Maintain* *Bill paying* *Portfolio mgt.* *Support: *Transfer problem* *General support* *Location of banks* *Products* *ATM/Online* *Credit card support* *Educational *Promotional	Interaction Mode *Banking transaction: *Transfer* *Inquiry* *Maintain* *Support: *Transfer problem* *General support* *Location of banks* *Products* *ATM/Online* *Credit card support*	Interaction Mode *Banking transaction: *Transfer* *Inquiry* *Maintain* *Deposit/ Withdrawal* *Portfolio* *Bank checks* *Travelers' check* *Support: *Transaction* *General* *Educational *Promotional
Information *Transactn: Date/ Time/c_ID/ Transaction Type/ Transaction related data	Information *Transaction: Date/c_ID/Transaction Type/data *Promotion/ Education: Date/c_ID/ promotionID/ Response *Support: Date/c_ID/ categories of questions - AQ/ live chat	Information *Transaction: c_ID/Time/Type/ data *Support: c_ID/Time/ Categories (Transaction vs. support)	Information *Transaction: c_ID/Time/Type/ data
*c_ID: customer ID			

havior and value in more detail, respectively. The attributes themselves were designed to mimic the database of a small retail bank.

<Table 2> displays a summary of the attributes that model the behavior of the bank customers. The theme of behavior is preference of interaction mode that is expressed by 'interaction mode used most often for transfer', 'interaction mode used most often for inquiry', and 'frequency of using each mode per month'. The value of a customer is expressed by a combination of savings account, checking account, and credit card balances.

Based on the attributes in Table 2, we developed total 35 attributes fields; 25 for behavioral aspect and 10 for value (see <Table 3> and <Table 4>).

Our data generator is designed to generate collections of customer populations based predetermined statistical characteristics of each population. We populated the database with nine discrete (i.e. non-overlapping) populations with different behavior attributes and distinguishing value propositions (see <Table 5>).

<Table 3> Behavioral attributes fields

Behavioral aspect					
1	Channel Banking	10	Banking portfolio mgt.	19	ATM bill paying
2	Channel ATM	11	Banking bank checks	20	Online transfer
3	Channel Online	12	Banking traveler checks	21	Online maintain
4	Channel Call center	13	ATM transfer	22	Online portfolio management
5	Banking transfer	14	ATM inquiry	23	Call center transfer
6	Banking inquiry	15	ATM maintain	24	Call center inquiry
7	Banking maintain	16	ATM withdrawal	25	Call center maintain
8	Banking withdrawal	17	ATM deposit		
9	Banking deposit	18	Online inquiry		

<Table 4> Value attributes fields

Value aspect			
1	Loan principal not available	6	Checking balance amount:
2	Loan principal amount:	7	Savings balance not available
3	Mortgage principal not available	8	Savings balance amount:
4	Mortgage principal amount:	9	Credit balance not available
5	Checking balance not available	10	Credit balance amount:

<Table 5>

		Behavior		
		A: Banking preferred	B: ATM preferred	C: Online & call center preferred
Value	High	population A_H	population B_H	population C_H
	Medium	population A_M	population B_M	population C_M
	Low	population A_L	population B_L	population C_L

With the combination of behavioral grouping and value grouping, the nine populations are labeled as follows:

A_L - population with Banking preferred, Low value

A_M - population with Banking preferred, Medium value

A_H - population with Banking preferred, High value

B_L - population with ATM preferred, Low value

B_M - population with ATM preferred,

<Table 6> Population: Behavioral Attributes

	Behavior A:	Behavior B:	Behavior C:

channel	banking = 1.0;
	ATM = 0.0;
	online = 0.0;
	call center = 0.0;
		banking = 0.0;
		ATM = 1.0;
		online = 0.0;
		call center = 0.0;
			banking = 0.0;
			ATM = 0.0;
			online = 0.5;
			call center = 0.5;

Mode banking	transfer = 0.5;
	inquiry = 0.5;
	maintain = 0.0;
	withdrawal = 0.0;
	deposit = 0.0;
	portfolio mgt = 0.0;
	bank checks = 0.0;
	traveler checks = 0.0;
		transfer = 0.0;
		inquiry = 0.0;
		maintain = 0.2;
		withdrawal = 0.2;
		deposit = 0.2;
		portfolio mgt = 0.4;
		bank checks = 0.0;
		traveler checks = 0.0;
			transfer = 0.0;
			inquiry = 0.0;
			maintain = 0.0;
			withdrawal = 0.0;
			deposit = 0.0;
			portfolio mgt = 0.0;
			bank checks = 0.5;
			traveler checks = 0.5;

ATM	transfer = 0.4;
	inquiry = 0.4;
	maintain = 0.2;
	withdrawal = 0.0;
	deposit = 0.0;
	bill paying = 0.0;
		transfer = 0.0;
		inquiry = 0.0;
		maintain = 0.0;
		withdrawal = 0.5;
		deposit = 0.5;
		bill paying = 0.0;
			transfer = 0.0;
			inquiry = 0.0;
			maintain = 0.0;
			withdrawal = 0.0;
			deposit = 0.0;
			bill paying = 1.0;

online	transfer = 1.0;
	inquiry = 0.0;
	maintain = 0.0;
	portfolio mgt = 0.0;
		transfer = 0.0;
		inquiry = 1.0;
		maintain = 0.0;
		portfolio mgt = 0.0;
			transfer = 0.0;
			inquiry = 0.0;
			maintain = 0.5;
			portfolio mgt = 0.5;

call center	transfer = 1.0;
	inquiry = 0.0;
	maintain = 0.0;
		transfer = 0.0;
		inquiry = 1.0;
		maintain = 0.0;
			transfer = 0.0;
			inquiry = 0.0;
			maintain = 1.0;

<Table 7> Population: Value Attributes

	Value L: low	Value M: medium	Value H: high
Loan principal amount: Mortgage principal amount: Checking balance amount: Savings balance amount: Credit balance amount:	$0 - $25,000 = 1.0 1 - 100,000 = 1.0 0 - 500 = 1.0 0 - 1,000 = 1.0 0 - 2,000 = 1.0	$25,001 - $50,000 = 1.0 100,001 - 500,000 = 1.0 501 - 5,000 = 1.0 1,001 - 30,000 = 1.0 2,001 - 10,000 = 1.0	$50,001 - $75,000 = 1.0 500,001 - 1,000,000 = 1.0 5,001 - 50,000 = 1.0 30,001 - 100,000 = 1.0 10,001 - 100,000 = 1.0

Medium value

B_H – population with ATM preferred, High value

C_L – population with Online & call center, Low value

C_M – population with Online & call center, Medium value

C_H – population with Online & call center, High value

Those nine discrete populations were generated by following statistical characteristics with 35 attributes (<Table 6> and <Table 7>). All of the attributes are encoded as binary attributes, 0 or 1 except for the monetary fields in Value. Therefore if we take a look at "Savings balance not available", "Savings balance amount", "Credit balance not available", and "Credit balance amount" in Value attributes fields, a record may look like: …… 1 0 0 2500. This part would read credit balance is not available, amount is set to 0, and credit balance is available, which is 2500 dollars.

At the end of the data fields, a label field is added to name each population for clear results. Therefore, a typical record in the database would look like:

1 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 1 0 0 35000 1 0 0 2000 0 3000 B_L

Due to the scaling issue of the monetary attributes, those fields were normalized by using the scaling formula, (actual entry - field min except for zero) / (field max – field min). The resulting database consisted of 9,000 records with 36 fields.

Given these *a priori* known populations, a successful methodology for campaign design should be able to positively identify a selected target population in the data. For our purposes here we consider the population B_H our target population. This target population represents individuals that use a mixture of bank tellers, ATMs, and some online banking and maintain large balances in their checking and savings accounts and also have loans and mortgages with this particular model bank. Here, the campaign design methodology should be able to clearly delineate the customers that are part of the B_H population as a single homogeneous group within the data.

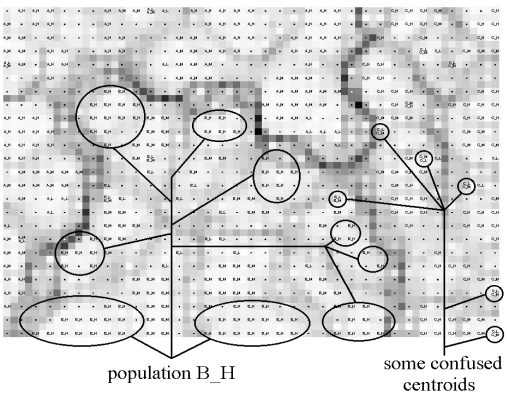
4.2 Campaign Design using a Single Self-Organizing Map

A common way to approach campaign design is to use a single SOM in order to find all the customers records that are similar to each other in the database. Using self-organizing maps has the advantage over algorithms such as k-means in that no *a priori* determination needs to be made on how many clusters are to be expected in the data. In our case, all the records for the population B_H should then lie in proximity to each other and should be identifiable as a cluster in the self-organizing map.

Our experiments, however, showed that even in our extremely idealized setting with discrete populations this approach failed to produce something intelligible. SOM-PAK (Kohonen et al., 1996) was used for analysis with initial radius 15,

the dimension of 30 by 23, and initial learning rate (alpha) 0.05. Iteration was set to 18000. [Figure 5] is a self-organizing map computed on our idealized customer database. Careful inspection shows that small clusters of our target population B_H appear throughout the map and it therefore becomes difficult to characterize what exactly B_H is and how to retrieve the records for this population from the database. This approach therefore fails our criterion as a successful campaign design tool. Our population B_H is not the only population scattered over the map, but other populations suffer from the same problem. This implies that the problem of population fragmentation is not the property of a single population but seems to be endemic to this single map approach. In addition to the population fragmentation we can also observe confusion in the map where a single centroid possesses labels from multiple different populations. Some of these confused centroids are indicated in [Figure 5].

One of the main parameters of self-organizing maps is the size of the map. It governs the ability of the map to form distinct clusters and minimize the corresponding quantization errors. However, in our case with the single map approach experiments have shown that increasing the map size alleviates the “confused centroid” issue but aggravates the population fragmentation into small clusters positioned randomly on the map. Convertsely, decreasing the map size alleviates population fragmentation but aggravates the “confused centroid” phenomenon. Therefore, we conclude that the single map approach has inherent problems that cannot be alleviated by fine-tuning the parameters of the self-organizing map.



[Figure 5] Self-Organizing Map Based on Our Idealized Customer Database

4.3 Campaign Design in Multiple Steps

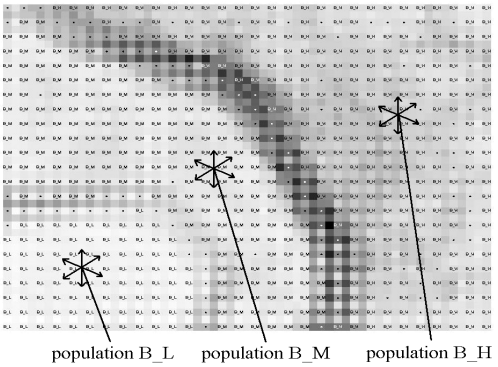
Our insight is that treating all the attributes in a data warehouse as the dimensions of a single Euclidean space simultaneously leads to confusion in the induced structures. If we treat groups of attributes as spanning separate Euclidean spaces and look for structure in each of these spaces separately we postulate that we will obtain much better results. In our idealized database we have two groups of attributes: behavioral attributes and value attributes. Here we employ the order for clustering “Behavior → Value”. Given this order, our methodology tells us that we need to first build a self-organizing map of the customer database based on the behavioral attributes. The populations were tested on the 25 behavioral fields only. The dimension was 30 by 23. The initial radius was set to 8, alpha to 0.05, and 9000 iterations were performed. Once this map is established we in turn

examine the emerging behavioral clusters for the value content. [Figure 6] shows the self-organizing map of our database based on the behavioral attributes. We can clearly see that the populations are being preserved according to the behavior which occurs in each population. Three major clusters can be seen to emerge from the figure, one for each behavioral group and each divided from the other by dark borders.

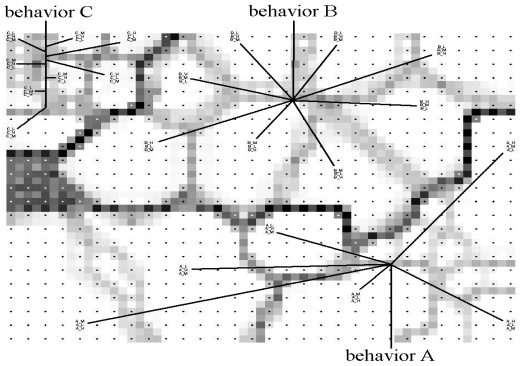
Here we are not interested in finding all possible populations but we are only interested in finding customers with behavior B. Therefore those group B records were tested with ten Value fields and a label field. The dimension was 30 by 23, with initial radius 15 and alpha 0.05. Iterations were 6000.

[Figure 7] shows the self-organizing map based on the value attributes constructed over the customers extracted from the previous map from the cluster that has behavior B.

We see that the top right corner of the map contains our target population B_H. We can also see that the complete population appears in that cluster, i.e., there is no population fragmentation. Furthermore, there is no confusion; none of the centroids has more than one label associated with it. Therefore, as one would expect in an example with discrete populations, we were able to identify and retrieve the target population completely.



[Figure 7] Value Clusters of Population with Behavior B



[Figure 6] Self-Organizing Map of Idealized Database Based on Behavioral Attributes

5. Discussion and Conclusions

Market segmentation for the identification of profitable customer sub-populations is an important strategy for marketers, particularly when considering marketing campaign design in the context of CRM. The starting point for any such strategy is a table that describes customers in terms of a set of attributes. In single SOM approaches this table is considered to define a multi-dimensional space and typically clustering algorithms are used to find structure in this space – to identify useful clusters

of customers. Contrary to intuition, this one step approach to clustering customers fails to fully identify important groups of customers. We have shown in our experiments that populations that are coherent by design are fragmented by these single SOM approaches, making a complete identification of these groups by the marketer virtually impossible. Furthermore, we found that these single SOM approaches tend to confuse populations as we showed with our experiments on discrete populations. An interesting observation in the case of self-organizing maps is that this problem cannot be alleviated by tuning map parameters, but this seems to be endemic to the single step approach. Our insight is that these shortcomings of the single SOM approaches seem to be due to the fact that such approaches ignore additional theme that is usually implicit within the attributes of the customer data. For example, typical customer databases do not describe the customer behavior using a single group of attributes but instead the behavior is described by several groups of attributes.

It seems that a multiple SOM approach can be used to identify target populations and the clusters without the same shortcomings as in the single SOM approaches. In order to overcome the shortcomings of a single SOM approach, advanced SOM approaches, such as hierarchical feature maps (Miikkulainen, 1990) and growing hierarchical SOM (Rauber, Merkl, and Dittenbach, 2002) were developed. However, hierarchical feature maps are focused on hierarchically structured data. As well, growing hierarchical SOM approach takes more time for the analysis because it uses all the attributes at the same time. In addition, due to its automatic detections of hierarchically added sub-structure, it is not easy to interpret the meanings of outcomes. On the other hand, this multiple SOM approach is more likely to deal with combined data with two or more groupings of attributes than hierarchically structured data, and since it is required to predefine the groupings and the order for clustering, it uses a part of attributes at a time. And therefore, it makes clustering time shorter. However, more experiments need to be conducted in order to confirm this observation since many other unanswered issues remain such as how to define the conceptual dimensions of dataset if not much are known, and how big the difference between the dimensions is required to be identified as dimension. As well, it will be valuable to examine multiple SOM approach and compare with K-means algorithm and other clustering methods in a statistical way. Therefore, future research should address those detailed issues and use real world data to see if the findings from computer-generated data are confirmed.

In sum, this study introduced possible shortcomings of single SOM approach as a preprocessor for clustering and raised several issues. This is important to recognize because if the one step approach does not reveal the hidden structure clearly even with the data set of the clean-cut structure, it would not work with more complicated data at all and based on confused SOM results, further clustering would not have much meaning. Therefore clustering with SOM as preprocessor will require more careful execution.

References

[1] Bang, J., N. Dholakia, L. Hamel, and R. R. Dholakia, "Data Mining Of CRM Knowledge Bases For Effective Market Segmentation", in: *Proc. ICEIS 2004, the 6th International Conference on Enterprise Information Systems* (Porto, Portugal, 2004).

[2] Berry, S. and K. Britney, "Market Segmentation: Key to Growth in Small Business Banking", *Bank Management* Vol.72(1996) 36～41.

[3] Curry, B., F. Davis, M. Evans, L. Moutinho, and P. Phillips, "The Kohonen self-organizing map as an alternative to cluster analysis: An application to direct marketing", *International Journal of Market Research* Vol.45(2003).

[4] Fayyad, U., Gregory Piatetsky-Shapiro, and Padhraic Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data," *Communications of the ACM* Vol.39 (1996) 27～34.

[5] Goldberg, D. E., Genetic algorithms in search, optimization, and machine learning. Reading, Mass.: Addison-Wesley Pub. Co., 1989.

[6] IDC, Worldwide CRM applications, market forecast and analysis summary 2001～2005, August, (2001).

[7] Kardes, F. R., " Consumer behaviour: managerial decision making', Addison-Wesley, Reading, MA 1999.

[8] Kiang, M. Y. and A. Kumar, "An evaluation of Self-Organizing Map Networks as a robust alternative to factor analysis in data mining applications", *Information systems Research* Vol.12, No.2(2001) 177.

[9] Kim, E., W. Kim, and Y. Lee, "Combination of multiple classifiers for the customer's purchase behavior prediction", *Decision Support Systems* Vol.34(2002), 167～175.

[10] Koh, H. C. and K. L. G. Chan, "Data Mining and Customer Relationship Marketing in the Banking Industry", *Singapore Management Review* Vol.24(2002), 1～27.

[11] Kohonen, T., Self-organizing maps, 3rd ed. Berlin; New York: Springer, 2001.

[12] Kohonen T., J. Hynninen J. Kangas and J. Laaksonen, (1996) "SOM_PAK : The Self-Organizing Map Program Package", *Technical Report A31*, Helsinki University of Technology, http://www. cis.hut.fi/nnrc/nnrc-programs.html.

[13] Kotler, P., Marketing Management : Analysis, Planning and Control. London: Prentice-Hall, 1997.

[14] Krieger, A. M. and P. E. Green, "Modifying Cluster-Based Segments to Enhance Agreement With an Exogeneous Response Variable", *Journal of Marketing Research* Vol.32(1996) 351～363.

[15] Kuo, R. J., K. Chang, and S. Y. Chien, "Integration of Self-Organizing Feature Maps and Genetic-Algorithm-Based Clustering Method for Market Segmentation", *Journal of Organizational Computing and Electronic Commerce* Vol.14(2004) 43～60.

[16] Kuo, R. J., L. M. Ho, and C. M. Hu, "Integration of self-organizing feature map and K-means algorithm for market segmentation", *Computers and Operational Research* Vol.29 (2002) 1475～1493.

[17] Libai, B., D. Narayandas, and C. Humby, "Toward and individual customer profitability model: A segment-based approach", *Journal of Service Research* Vol.5(2002), 69.

[18] Luxton, R., "Marketing campaign systems-the secret to life-long customer loyalty", *Journal of Database Marketing* Vol.9 (2002), 248～258.

[19] Mackinnon, M. J. and Ned Glick, "Data mining and knowledge discovery in databases - an overview", *Australia & New Zealand Journal of Statistics* Vol.41(1999), 255～275.

[20] Mazanec, J. A., Exploratory market structure analysis: topology-sensitive methodology, Vienna University of Economics and Business Administration Vol.31(1999).

[21] Miikkulainen, Risto, "Script recognition with hierarchical feature maps", *Connection Science* Vol.2, No.1, 2(1990), 83～101.

[22] Mitchell, S., "The new age of direct marketing", *Journal of Database Marketing* Vol.10(2003) 219.

[23] Pampalk, Elias, Gerhard Widmer, and Alvin Chan (2003), "A New Approach to Hierarchical Clustering and Structuring of Data with Self-Organizing Maps," *Technical Report OeFAI-TR-2003-09*.

[24] Papatla, P. and A. Bhatnagar, "Shopping Style Segmentation of Consumers", *Marketing Letters* Vol.13(2002), 91.

[25] Rauber, Andreas, Dieter Merkl, and Michael Dittenbach, "The Growing Hierarchical Self-Organizing Map : Exploratory Analysis of High-Dimensional Data," *IEEE Transaction on Neural Networks* Vol.13, No.6(2002), 1331-1341.

[26] Reichheld, F. F. and W. E. Sasser, "Zero Defections Quality comes to Services", *Harvard Business Review* Sept/Oct (1990), 301～307.

[27] Ryals, L., "Creating profitable customers through the magic of data mining", *Journal of Targeting, Measurement and Analysis for Marketing* Vol.11(2003), 343～349.

[28] Schewe, C. D. and G Meredith, "Segmenting global markets by generational cohorts: Determining Motivations by Age", *Journal of Consumer Behavior* Vol.4, No.1(2004), 51～63.

[29] Shaw, M. J., C. Subramaniam, G. W. Tan, and M. E. Welge, "Knowledge management and datamining for marketing", *Decision Support Systems* Vol.31(2002).

[30] Smith, K. A. and A. Ng, "Web page clustering using a Self-Organizing Map or user navigation patterns", *Decision Support Systems* Vol.35, No.2(2003), 245.

[31] StatSoft, Electronic Statistics Textbook. Tulsa, OK: StatSoft, Inc., 2004.

[32] Walsh, G., T. Henning-Thurau, V. Wayne-Mitchell, and K. P. Wiedmann, "Consumers' decision-making style as a basis for market segmentation", Journal of Targeting, Measurement and Analysis for Marketing Vol.10(2001), 117.

[33] Witten, I. and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. 1999 : Morgan Kaufman

요약

고객관계관리의 시장 세분화를 위한 Self-Organizing Maps 재고찰

방정혜* · Lutz Hamel** · Brian Ioerger***

본 논문은 고객관계관리를 위한 시장 세분화를 하기 위해 자주 사용되는 SOM에 대하여 고찰한다. 일반적으로, SOM은 군집의 수를 미리 파악하기 위하여, 구체적인 군집 분석이 이루어지기 이전에 사용된다. 그러나 인터넷이 발달하고 수집 가능한 데이터의 종류와 양이 증가함에 따라 복합적인 분석이 필요하게 되었다. 또한, 그에 따라 한가지 주제만으로 군집을 파악하는 것보다는 여러 가지의 주제들을 대상으로 고객데이터의 군집을 파악해야 하는 경우가 많이 발생하게 된 것이다. 따라서 이 논문에서는 이렇게 한가지의 주제가 아닌 여러 가지의 주제로 군집분석을 할 경우 한번으로 이루어지는 SOM 어프로치가 과연 군집의 수를 파악할 수 있는지를 실험하였다. 이미 구조를 알고 있는 데이터를 생성하여 실험을 해본 결과, 전체 데이터를 대상으로 여러 주제를 한꺼번에 포함시킨 경우 (single SOM 방식) 에는 그 구조를 제대로 파악하지 못하였으며, 하나의 주제마다 각기 다른 SOM을 사용(multiple SOM 방식)한 결과, 미리 정해졌던 구조를 제대로 파악할 수 있었다. 따라서 이 논문은 군집분석을 하게 될 경우, 좀더 조심스러운 접근법이 필요하며, 여러가지 주제를 포함하고 있는 데이타를 다룰 경우, SOM 분석 방법에 대하여 논의하였다.

* 국민대학교 경영대학
** Department of Computer Science and Statistics, University of Rhode Island, USA
*** Department of Computer Science and Statistics, University of Rhode Island, USA