
Original Article

Data mining framework for customer lifetime value-based segmentation

Received (in revised form): 6th November 2011

Harsha Aeron

is an Analyst at Latent View Analytics. He acquired his Doctoral Degree in IT and systems from the Indian Institute of Management, Lucknow, India. His main research interests are business intelligence, data mining applications, customer relationship management and customer lifetime value. He has published in *Journal of Database Marketing & Customer Strategy Management*, and *International Journal of Business Information Systems*.

Ashwani Kumar

is presently an Associate Professor in IT and systems at the Indian Institute of Management, Lucknow. His current research interests are business intelligence, data mining and its applications in business, especially marketing and finance. He received his PhD from the ABV-Indian Institute of Information Technology and Management, Gwalior; his MS from National University of Singapore; his MBA from the University of Melbourne, Australia; and his BTech (EE) from the Indian Institute of Technology, Kanpur. He has published in journals such as *Journal of Database Marketing & Customer Strategy Management*, *Intelligent Data Analysis*, *International Journal of Information Technology and Decision Making*, *International Journal of Intelligent Systems in Accounting, Finance and Management* and so on.

Janakiraman Moorthy

is a Professor in the Indian Institute of Management, Calcutta. He specializes in advanced marketing research, new product development, customer value creation and market orientation of firms. His recent publications were in *Journal of Database Marketing & Customer Strategy Management* and *Marketing Science*. His current work is focused on reviewing the methodologies for customer valuation and marketing productivity analysis. He received his PhD from the Indian Institute of Management, Ahmedabad, India, and was Global Research and Project Director of the Institute for Customer Relationship Management, Atlanta, USA.

ABSTRACT Estimating Customer Lifetime Value (CLV) is essential for firms competing in data-rich environments. Segmentation on the basis of CLV is helpful in customization of products and services by justification of resource allocation. Model-based automated decision making is likely to penetrate various marketing decision-making environments. We are presenting a framework for customer lifetime value-based segmentation. The framework automates two decisions: first, selection of variables; and second creation of optimal segments on the basis of CLV. The framework uses clustering for segmentation and genetic algorithm for optimization.

Journal of Database Marketing & Customer Strategy Management (2012) 19, 17–30. doi:10.1057/dbm.2012.1; published online 6 February 2012

Keywords: database marketing; customer lifetime value; segmentation; Recency-Frequency-Monetary Value (RFM); genetic algorithm; clustering

Correspondence:

Harsha Aeron
House No. 8, Nav Vihar
Colony, Samad Road,
Aligarh – 202001,
Uttar Pradesh, India
E-mail: harsha.aeron@gmail
.com

INTRODUCTION

In the last few decades, digital revolution has led to a flood of data in the firm's databases. The data can help a firm, through Knowledge Discovery in Databases

(KDD), in taking more informed decisions, thus making it more efficient, less intuitive and faster. KDD is the process of finding useful information and patterns in data, whereas Data Mining (DM) is the use of

algorithms to extract the information and patterns derived by the KDD process. DM is defined as the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.¹ It leads to a better understanding of the data collected, as a result of the interaction of the firm with its environment and internal processes. The interaction with the environment means dealing with the customers, suppliers and competitors, whereas internal processes means tracking sales, managing product portfolio and so on.

Increasingly, companies are changing their approach from product centric to customer centric. In the former approach, products are assets and the company's focus is to sell more and more products. However, products and cost advantages can be replicated by the competition. On the other hand, in the customer-centric approach, the company considers customers as assets and focuses on acquiring and retaining more customers. Customer base can be seen as an asset that has a value and can help in deciding acquisition and retention.² The customer base can be a source of competitive advantage for the company. In the recent times, more and more companies are becoming service companies, thus enhancing the importance of the customer-centric approach.

The customer-centric approach is further strengthened by the advent of relationship marketing in the last few decades. Relationship marketing thrives on the fact that retaining and selling to the existing customers is much more profitable than acquiring new customers. The advent of relationship marketing along with technological advancements has led to Customer Relationship Management (CRM). To implement CRM effectively, a firm must be able to form a learning relationship with customers through

a four-step approach: (1) *Notice* – what its customers are doing; (2) *Remember* – what the firm and its customers have done over time; (3) *Learn* – from what it has remembered; and finally (4) *Act* – on what it has learned to make customers more profitable. The learning part of CRM implementation is done with the help of DM.

Relationship marketing is costly, as making and maintaining relations requires allocation of resources. It might not be feasible to keep good relations with all customers at all times. As marketing is the art of attracting and keeping profitable customers,³ it seems logical to choose your customers properly for both acquisition and for retention. One way of doing so can be through Customer Lifetime Value (CLV). A profitable customer is defined as 'a person, household, or company whose revenues over time exceed, by an acceptable amount, the company cost of attracting, selling, and servicing that customer'.³ The excess is called CLV. CLV is also known by other names such as customer profitability, lifetime value (LTV), customer equity and so on.

CLV models can be useful in identifying a firm's customer characteristics, whereas segmentation of customers on the basis of these characteristics/variables may help in resource allocation for different marketing activities. CLV is important for companies having their business models based on the Internet, for example Google. As these companies do not have any fixed asset but only intangible assets, the value of their customers may help in valuing these companies.⁴

Different benefits of CLV are as follows:⁵

- Focus on CLV helps in deciding on what data to collect at the operational level, as there is no dearth of information available today. This information should be turned to intelligence to support decision making.

- It highlights the value of different assets for the firm. The non-relational assets are valuable to the extent that they increase the value of customer assets.
- The role of the marketing department is redefined so that it manages customer assets in a way that facilitates customer equity.
- At the organizational level, it provides metrics to estimate the effect of various organizational activities on customer equity.
- It helps in developing more value-based segmentation of customers.

This article focuses on segmentation on the basis of CLV.

Firms have long recognized the importance of market segmentation for tailoring their marketing mixes according to the specific needs and preferences of different customer groups. 'Market segmentation involves viewing a heterogeneous market as a number of smaller homogenous markets in response to differing preferences, attributable to the desires of customers for more precise satisfaction of their varying wants'.⁶

Those firms that master information about their customers will thrive by delivering customized, highly valued offerings, whereas those that do not will be at a competitive disadvantage.⁷ Market segmentation helps in mastering the customer information, as it offers four major benefits: (1) designing responsive products to meet the needs of the target segment; (2) determining effective and cost-efficient promotional strategies; (3) evaluating market competition, in particular the company's market position; and (4) providing insights into particular marketing strategies.⁸ As companies are becoming more and more customer centric, many of the fundamental marketing assumptions are challenged and thus is the basis of segmenting customers.⁹

Market-oriented businesses have two main goals: (1) to satisfy the needs of the customer by offering products that meet the desires, requests and demands of the customers; (2) to satisfy the needs of the business units by carrying out exchanges that result in long-term profitability.¹⁰ The segmentation on the basis of commonly deployed criteria such as geographic, demographic, psychographic and behavioristic may serve only the first purpose, whereas segmentation on the basis of profitability serves both purposes at the same time. The effect of different marketing actions can be estimated both by a possible change in segment profitability and switching probabilities of customers from one segment to another. It enables companies to estimate long-run customer equity profitability using switching probabilities and can be evaluated on the basis of the company's history with the segments.

The authors discuss the long-run future of decision support systems in marketing. They argue that a growing proportion of marketing decisions can not only be supported, but may also be automated.¹¹ From the standpoint of efficiency (for example, management productivity) and effectiveness (for example, resource allocation decisions), such automation is highly desirable. The authors describe how model-based automated decision making is likely to penetrate various marketing decision-making environments.

The article proposes a DM framework for CLV framework. The framework uses genetic algorithm-based clustering. The framework automates two decisions: first, it automatically segments customers on the basis of CLV, making as distinct segments as possible. The framework decides the number of clusters. Second, the framework chooses among the set of input variables the most important variables to give distinct customer segments.

LITERATURE REVIEW

On the basis of the review of CLV metric papers,^{4,12} the metrics can be divided into two broad classes as follows:

- Metrics calculating the total value contributed by a particular customer or a segment of customers; this classification can be further sub-classified as metrics for individual customers and metrics for customer portfolio.
- Metrics calculating the value contributed by a customer at the time of acquisition, retention or expansion.

One set of metrics^{13,14} have calculated CLV for individual customers. Few metrics take into account past transactions and inactivity of the customer.^{14,15} The other set of CLV metrics are given for a cohort of customers.^{12,16} These help in analyzing the effects of elements of marketing on the long-term value of the firm's customer base, to calculate profitability from a growing customer base and to give a metric to gauge revenue from a cohort of customers at different relationship stages. A common drawback with these metrics is that the customer heterogeneity is not represented in the metrics.

The other set of CLV metrics is one in which value from customer acquisition, customer retention and customer expansion (cross-selling or margin) is calculated.^{2,17,18} In these metrics, either a component of CLV such as acquisition or retention is represented, or a link between two components is shown. The main drawback with them is that they cater to only a portion of CLV, and not CLV of a customer as a whole. Thus, the value of a customer, throughout her lifetime, is not captured. In addition to the aforementioned drawbacks of each set of metrics, almost all metrics are given for consumer products and not for services where customer acquisition and retention are in a meaningful context.

CLV-based segmentation can be done in different ways. Four generic ways to segment customer bases are reported:⁹ (1) segmentation based on combining relationship revenue and relationship cost; (2) segmentation based on relationship volume; (3) segmentation based on customer relationship profitability; and (4) segmentation based on combining relationship volume and customer relationship profitability. A possible segmentation approach commonly known as a customer pyramid divides the customer base into tiers, based on the CLV^{19,20} and way to move customers between segments.²⁰

Customer segmentation methods using CLV can be classified into three categories as follows:

1. Segmentation by using only LTV values.
2. Segmentation by using LTV components.
3. Segmentation by considering both CLV values and other information.

In the first method, customers are arranged in the descending order of CLV (Figure 1). In this case, only CLV values are used for segmenting customers. The list is divided by its percentile.

In the second method, segmentation is performed on the basis of components of

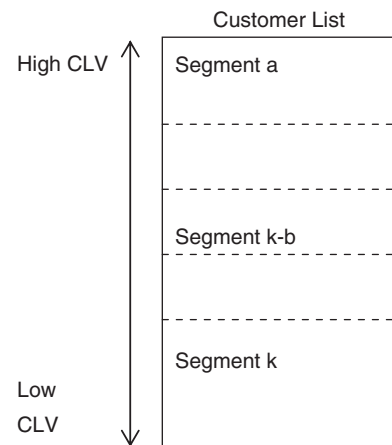


Figure 1: Segmentation by CLV values only.

CLV²¹ (Figure 2). CLV is divided into three components: current value, present value and customer loyalty. Current value provides financial viewpoints and potential value indicates cross-selling opportunities. Customer loyalty estimates durability of the previous two values.

In the last method, CLV is used along with other information such as socio-demographic information and transaction history to segment customers (Figure 3).

Various DM techniques have been used for segmentation. One of the most common techniques used is clustering. The usual objective of a clustering analysis is to separate objects or a respondent into groups such that homogeneity is maximized within the groups and heterogeneity is maximized within the groups. Cluster analysis is found to be an effective segmentation technique.²² Clustering has been used extensively for

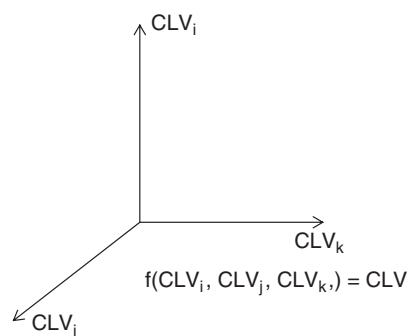


Figure 2: Segmentation by CLV and CLV components.

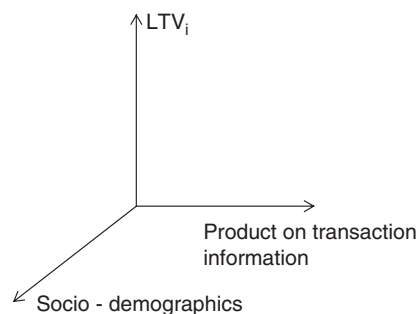


Figure 3: Segmentation by CLV and other variables.

segmentation either separately or in a hybrid form along with other DM techniques.^{23–26} The second technique, which has found an early application for segmentation, is decision trees in the form of AID²⁷ and CHAID.²⁸ The main advantage of decision trees is their explainability. Recently, logistic regression²⁹ and artificial neural networks³⁰ have also found their application for market segmentation. Most of the DM techniques have been applied to segment on the basis of demographics, psychographics and behavioristic variables. We have used genetic algorithm-based clustering to segment on the basis of customer CLV.

GENETIC ALGORITHM-BASED CLUSTERING

Genetic algorithm-based clustering literature is divided broadly into two categories; that is, for a fixed number of clusters and second for a variable number of clusters.³¹ We will segregate the literature under each heading by encoding schemes, operators such as crossover, mutation, selection and fitness function.

The fixed number of clusters K is valid in a problem where the operator has some domain knowledge and thus can provide the number of clusters or the expected range of clusters for the solution space.

Encoding schemes

Encoding schemes can be broadly classified into three categories: binary, integer and real. In the binary encoding scheme, the representation can be in two forms. One form of representation of a clustering solution is in the form of a binary string of length N , where N is the number of data set objects. There are other ways of representing a clustering solution using binary encoding. One way is matrix-based binary encoding. In this method, a $K \times n$ matrix is used for representation, in which the rows represent clusters and

columns represent objects. In integer-based encoding, integer values are used to represent the clusters in the chromosome.

Crossover

A uniform crossover is adopted³² in which the parent genotypes swap their *i*th bits. A context-sensitive operator that is based on exchanging medoids from the parents is proposed.³³ The set of medoids encoded into the offspring is based iteratively, adding one medoid at a time, until *k* different medoids have been represented. In pairwise crossover, clusters encoded into different parents are paired according to the similarities of their centroids. An offspring is then generated by randomly taking one centroid from each pair of clusters.

Mutation

Different kinds of mutation operators are used; Murthy and Chowdhary³⁴ used a mutation operator that randomly changes the gene value (cluster label) of some randomly selected variable,³⁴ a mutation operator that changes a gene value depending on the distances of the cluster centroids from the corresponding object.³⁵ In particular, the probability of changing a gene value to a given cluster label is higher if the centroid of the corresponding cluster is closer to the object.

Selection

The most popular selection is proportional selection followed by elitist selection.

Fitness function

The different fitness functions used are: the ratio of the distance between centroids and average intra-cluster distances as the fitness function,³⁶ the sum of within-cluster distances as fitness function³⁷ and the sum of squared Euclidean distances as fitness function.³⁸

Most of the encoding schemes used in evolutionary algorithms are capable of

estimating the number of clusters (*k*) and are similar to the encoding scheme employed in evolutionary algorithms for which *k* is assumed to be known or set *a priori*. An evolutionary algorithm for clustering, named Evo Cluster,³⁹ encodes a partition in such a way that each gene represents one cluster and contains the label of the objects grouped in it.

A slightly different encoding scheme allows working with constant length genotypes.⁴⁰ Basically, the number of clusters encoded by a given genotype is assumed to be in the range of [*k*-min, *k*-max] – where *k*-min and *k*-max are the minimum and maximum number of clusters, respectively – and a ‘don’t care’ symbol (#) is used to fill in genotypes whose *k* is less than *k*-max.

Crossover

Different crossover approaches covered are: a uniform crossover in the early generation and a two-point crossover in later generation;⁴¹ a two-point crossover that allows exchanging real-valued cluster prototypes from a pair of genotypes;⁴⁰ a one-point crossover that manipulates the edges of a decision tree, in which the nodes represent the data set objects and the edges correspond to the proximity index between them.⁴²

Mutation

Cole⁴³ uses three mutation operators designed for integer encoding. The split cluster-oriented operator probabilistically selects a cluster from a particular partition and moves objects from that cluster into a new cluster. The mega cluster-oriented operator moves all the objects from one cluster to another pre-existing cluster. Finally, the move object-oriented operator shifts objects between clusters already encoded on a genotype. Other mutation operators used are: two cluster-oriented operators⁴⁴ and equation-based mutation for real encoding.⁴⁰

Selection

In selection again proportional and elitist are used for selection.

Fitness function

Different indexes such as Tseng and Yang Index,⁴⁵ Kaufman and Rousseeuw Index,⁴⁶ Davies–Bouldin Index,⁴⁷ Calinski and Harabasz Index⁴⁸ have been used. Tseng and Yang Index measures the difference between inter-distance of clusters and intra-distance of clusters. Kaufman and Rousseeuw Index measures the dissimilarity between two nearby clusters and tries to reduce that. Davies and Bouldin Index measures intra-cluster distance to inter-cluster distance; finally, Calinski and Harabasz Index measures between-cluster and pooled within-cluster sum of squares matrices.

METHODOLOGY

The methodology used for the framework mainly consists of the following steps:

- Deciding the appropriate CLV metric according to the context.
- Deciding the appropriate validation metric.
- Designing the structure of chromosome for Genetic Algorithm.
- Defining the fitness function for Genetic Algorithm.

The framework of CLV-based segmentation is given in Figure 4.

The framework starts with designing the structure of the chromosome, then deciding the initial population and defining the fitness function. CLV values of each customer are estimated. Customers are segmented on the basis of clustering (fuzzy

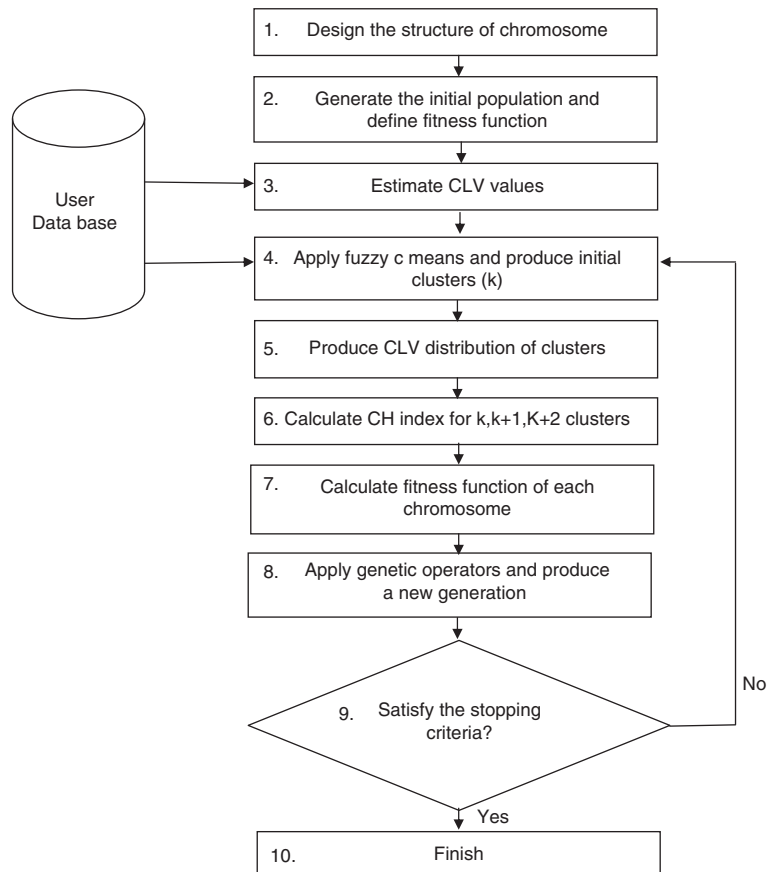


Figure 4: Framework for CLV-based segmentation.

clustering). Once CLV values and segments are obtained, then CLV distributions of each segment are produced. The validity index (CH index) is applied to the CLV distributions for $k, k+1, k+2$ clusters. Fitness function of each chromosome is calculated. Genetic operators, such as crossover, mutation and so on, are applied and a new generation is created. The process is repeated till one of the convergence criteria is achieved. The novelty in the framework is that validity metric is applied on CLV distribution rather than segment elements.

Structure of chromosome

The chromosome is binary coded and has 30 genes. Twenty-five genes represent factors of the data set and five genes are for the clustering solution. Binary encoding is used for the ease of using it.

Generate the initial population and define fitness function

In this step, the system generates initial population that would be used to find global optimum. Each chromosome has 30 genes. The first 25 genes represent the number of variables in the data set. All of these or a subset of these variables would be used for segmentation. A particular variable is used if the corresponding value of the allele is 1; otherwise, it is not included in the segmentation. In the initial population, different subsets of variables are included for segmentation, which keeps on changing as the GA iteration increases with the assumption of giving better subsets for segmentation. The last five genes represent the number of clusters. Thus, a value of 00011 in the last five genes represent three clusters would be formed for testing further.

For defining the fitness function, we use the CH index in a different way. The

CH index for n data and K clusters is computed as

$$[\text{trace } B / (K - 1)] / [\text{trace } W / (n - k)]$$

Here, B and W are the between- and within-cluster scatter matrices. The maximum hierarchy level is used to indicate the correct number of partitions in the data. The trace of the between-cluster scatter matrix B can be written as

$$\text{trace } B = \sum_{k=1}^K n_k \|z_k - z\|^2 \quad (1)$$

where n_k is the number of points in cluster k and z is the centroid of the entire data set.

The trace of the within-cluster scatter matrix W can be written as

$$\text{trace } W = \sum_{k=1}^K \sum_{i=1}^{n_k} \|x_i - z_k\|^2 \quad (2)$$

The CH index is applied on the CLV distributions of different segments as compared with other techniques where the index has been applied to segments.

Genetic algorithm starts with a random number of clusters. The values CH index is calculated at this cluster number and the consecutive two more clusters numbers. The change of the value of CH index is calculated from one cluster to another. Further, the change in the change of the CH value is calculated. The fitness function is to maximize this change in the change of CH index.

CLV estimation

The following formula is used to compute CLV.

$$\text{CLV} = \text{margin} \times \text{revenue/transaction} \times \text{DET} \quad (3)$$

where discounted expected transactions (DET) is derived from the Pareto/NBD

Table 1: Value of purchase parameters

S. no.	<i>r</i>	<i>Alpha</i>	<i>Beta</i>	<i>S</i>
1	0.9854	24.6564	0.9729	4.566
2	0.9972	24.6098	0.9757	4.5549
3	0.996	24.8314	0.9931	4.5601
4	0.9888	24.9516	0.8586	4.8635
5	0.9961	24.9365	1.0241	4.4661
6	1.0029	24.4082	0.8276	4.9338
7	0.9934	24.4855	0.902	4.7045
8	0.9854	25.0402	0.9981	4.5418
9	0.982	25.0024	0.8411	4.906
10	1.0029	24.6175	0.8132	5.0129
Mean	0.99301	24.75395	0.92064	4.71096

model.¹⁴ The only customer-level information that this model requires is recency and frequency. The notion used to represent this information is $(X=x, t_x, T)$, where x is the number of transactions observed in the time interval $(0, T)$, and t_x ($0 < t_x \leq T$) is the time of the last transaction. In other words, recency and frequency are sufficient statistics for an individual customer’s purchasing history.

Drawing on standard representations, we could compute the number of DET for a customer with observed behavior $(X=x, t_x, T)$ as

$$DET = \sum_{i=1}^n (E[Y(t) | X = c, t_x, T] - E[Y(t-1) | X = x, t_x, T]) / (1 = d)^t \tag{4}$$

The above formula is used as follows in our algorithm

$$DET(\delta | r, \alpha, s, \beta, X = x, t_x, T) = \frac{\alpha^r \beta^s \delta^{s-1} \Gamma(r+x+1) \psi[s, s; \delta(\beta+T)]}{\Gamma(r)(\alpha+T)^{r+x+1} L(r, \alpha, s, \beta | X = x, t_x, T)} \tag{5}$$

where r, α, s and β are the Pareto/NBD parameters; $\psi(\cdot)$ is the confluent hypergeometric function of the second kind; and $L(\cdot)$ is the Pareto/NBD likelihood function.⁴⁹

For determining the parameters of algorithm r, α, s, β , we have determined the parameters of purchase rate and death rate. Purchase rate distribution is determined as the rate of purchase at which donation is happening and death rate is determined by the distribution of inter-purchase interval. We have taken 10 samples of our data where at least two transactions have happened so as to avoid null values in the inter-purchase table. The values of purchase parameters are given in Table 1.

We have used these mean values of parameters in our calculations.

Producing clusters and CLV distributions

For segmenting the database, fuzzy c-means clustering is used. A CLV distribution is created by dividing the range between the minimum and maximum values of CLV in 10 equal bins and counting the number of CLV values in each bin. We have found the CLV distribution to be highly skewed towards the left, that is, there are few customers with very high CLV values.

Segment optimization

Once the CLV distribution of each cluster is obtained, the CH value is calculated for those distributions. Similarly CH value is calculated for $k+1$ and $k+2$ numbers of clusters is calculated and fitness function is evaluated for each chromosome.

Genetic operators such as crossover, mutation and selection are applied to produce a new generation of chromosome. The process continues till one of the stopping criteria is reached.

RESULTS AND ANALYSIS

The data we have used are Direct Marketing Education Foundation data set 1 distributed by Teradata Research Center, USA. The data are of a non-profit organization that uses direct mail to solicit additional contributions from past donors. There is one record per past donor and a total of 99200 records. Each record contains 77 fields. Data contain the latest 1–10 donations by date, dollar amount and solicitation source, the latest 1–11 solicitations by date and type, some lifetime elements of the relationship and minimal demographics. Data are from October 1986 to June 1995.

Data treatment and description

All dates were converted into time measuring backwards from June 1995, that is December 1994 is measured as 6 months. This is done to make data compatible with the requirements of the Pareto/NBD model. We have mainly used three sets of data variables: first, those that are used to calculate CLV values according to the Pareto/NBD model; second is recency, frequency and monetary values of contributions; and third is those variables that may be useful in segmentation such as the amount contributed over a lifetime, number of times solicited and so. We have mainly done away with variables describing the solicitation pattern, product codes of donation and solicitation, and some other variables such as gender, club membership and so on. The reason for non-inclusion of these variables is mainly the ease of the algorithm. The total columns included in the data analysis and their description is given in the Appendix.

The data are normalized and all records with missing values are removed. A sample

of 10000 records is randomly selected by using SAS command. Once the records with outliers are removed from the data set, the number of records reduces to less than 10000.

Framework parameter options

Our framework has different components such as CLV metric, type of clustering algorithm used, validation metric used, design of fitness function and values of Genetic Algorithm parameters. We got the best results by using fuzzy c means clustering, Calinski and Harabasz metric as validation metric, change in metric value as fitness function, scaling function as rank, selection as remainder, mutation as Gaussian and crossover as Two-Point. Various other options were also tried but to get only inferior results as compared with the reported results. The GA plot for the best solution is given below. The chromosome for the best plot is taken as a solution as the plot has stabilized (Figure 5).

The outputs of the framework for other combination of framework parameters are either not stabilizing or are stabilizing at two clusters, which is the minimum number of clusters given in the framework.

The chromosome for the best solution is given in Figure 6.

Whereas the last five alleles represent the cluster solution, the first 25 alleles represent the variables selected to give the optimum clusters. The number of clusters obtained is six, which means the cluster solution is seven clusters as the solution is the knee point in the fitness function curve, whereas the output in the form of chromosome is the point just before the knee point (Figure 7).

The attributes selected are as follows:

1. largest contributed amount;
2. amount contributed over a lifetime;
3. latest contribution amount;
4. 3rd latest contribution amount;

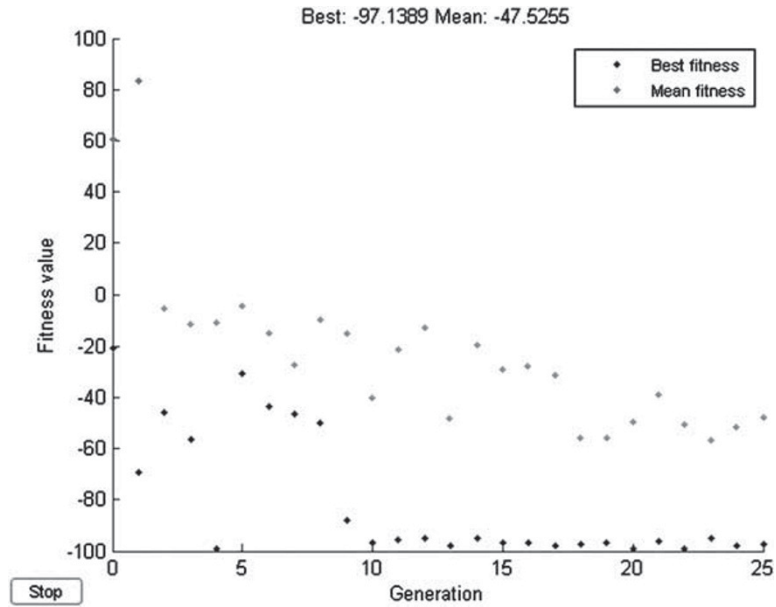


Figure 5: GA solution depicting the best plot.

F	T	F	T	F	T	F	T	F	F	F	T	F	F	F	F	T	F	F	T	T	T	F	F	T	T	F
A	R	A	R	A	R	A	A	A	A	R	A	A	A	A	R	A	A	R	R	R	R	A	A	R	R	A
L	U	L	U	L	U	L	L	L	U	L	L	U	L	L	U	L	L	U	U	U	U	L	L	U	U	L
S	E	S	E	S	E	S	S	S	E	S	S	E	S	S	S	E	S	S	E	E	E	S	S	E	E	S
E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E

Figure 6: Chromosome for best solution.

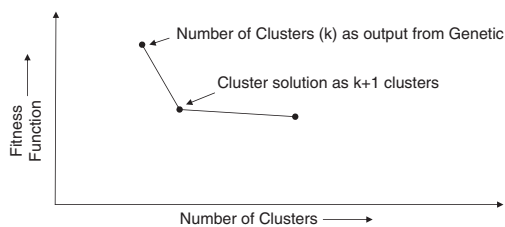


Figure 7: Knee solution.

- 5. 7th latest contribution amount;
- 6. 10th latest contribution amount;
- 7. 4th latest contribution time;
- 8. 7th latest contribution time;
- 9. 8th latest contribution time;
- 10. 9th latest contribution time;
- 11. 10th latest contribution time.

The attributes selected are having good explanatory value. Among the RFM (recency, frequency and monetary) variables, all three type of relevant variables are selected. Among the amount variables, the largest contributed amount by a donor and the amount contributed over a lifetime are the most distinguishing variables. For recency and amount variables, the amounts contributed at the third, seventh and last month are important. The results are convincing as the duration at which the amount is considered is approximately 3 months. For the frequency variables, a pattern is selected towards the beginning of the observation window and then an observation is selected at a duration of

3 months. Thus, the algorithm has selected a more permanent pattern to distinguish among customers.

CONCLUSION

Estimating CLV has many advantages, and one of the advantages is segregation of customers. Segregating customers on the basis of CLV helps in policy customization and resource allocation towards customers. Customer segmentation has been tried using demographic variables, purchase-based variables and so on. One more approach of customer segmentation can be on the basis of CLV. Different DM techniques, such as clustering, neural networks, genetic algorithms and so on, have been applied for segmentation. However, research where DM techniques are applied for customer segmentation on the basis of CLV is lacking.

We have proposed a DM framework for doing CLV-based segmentation. The framework is based on Genetic Algorithm-based clustering. The framework is novel in two ways; first, it automatically segments customers on the basis of CLV so that the segments are as distinct as possible. The framework decides the number of clusters. Second, the framework chooses among the set of input variables the most important variables to give distinct customer segments. The variables selected have good interpretability and explanatory value, which is very useful for decision making and application in customer segmentation and CRM. The DM framework can be made more comprehensive by estimating CLV value, incorporating all CLV parameters and estimating/predicting these CLV parameters by DM techniques as showed in Aeron et al.⁵⁰

REFERENCES

- 1 Hand, D., Mannila, H. and Smyth, P. (2005) *Principles of Data Mining*. New Delhi: Prentice Hall of India Private Limited.
- 2 Blattberg, R. and Deighton, J. (1996) Manage marketing by the customer equity. *Harvard Business Review* 74(4): 133–144.
- 3 Kotler, P. and Armstrong, G. (1996) *Principles of Marketing*, 7th edn. Englewood Cliffs, NJ: Prentice-Hall.
- 4 Jain, D. and Singh, S.S. (2002) Customer lifetime value research in marketing: A review and future directions. *Journal of Interactive Marketing* 16(2): 34–46.
- 5 Hogan, J.E., Lemon Katherine, N. and Rust Roland, T. (2002) Customer equity management: Charting new directions for the future of marketing. *Journal of Service Research* 5(1): 4–12.
- 6 Smith, W. (1956) Product differentiation and market segmentation as alternative marketing strategies. *Journal of Marketing* 21(1): 3–8.
- 7 Shapiro, C. and Varias Hal, R. (1999) *Information Rules: A Strategic Guide to the Network Economy*. Boston, MA: Harvard Business School Press.
- 8 Weisntein, A. (1994) *Market Segmentation: Using demographics, Psychographics and other Niche Marketing Techniques to Predict and Model Customer Behavior*. Clinton, Chicago, IL: Probus Publishing Company.
- 9 Storbacka, K. (1997) Segmentation based on customer profitability – Retrospective analysis of retail bank customer bases. *Journal of Marketing Management* 13(5): 479–492.
- 10 Helgesen, O. (2005) Customer segments based on customer account profitability. *Journal of Targeting, Measurement and Analysis for Marketing* 14(3): 225–237.
- 11 Bucklin, R.E., Lehmann, D.R. and Little, J.D.C. (1998) From decision support to decision automation: A 2020 vision. *Marketing Letters* 9(3): 234–246.
- 12 Gupta, S. et al (2006) Modelling customer lifetime value. *Journal of Service Research* 9(2): 139–155.
- 13 Berger, P.D. and Nasr, N.I. (1998) Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing* 12(1): 17–29.
- 14 Schmittlein, D.C., Morrison, D.G. and Colombo, R. (1987) Counting your customers: Who are they and what will they do next? *Management Science* 33(1): 1–24.
- 15 Aeron, H., Bhaskar, T., Sundararajan, R., Kumar, A. and Janakiraman, M. (2008) A metric for customer lifetime value of credit card customers. *Journal of Database Marketing & Customer Strategy Management* 15(3): 153–168.
- 16 Johnson, M.D. and Selnes, F. (2004) Customer portfolio management: Toward a dynamic theory of exchange relationships. *Journal of Marketing* 68(2): 1–17.
- 17 Hansotia, B. and Paul, W. (1997) Analytical challenges in customer acquisition. *Journal of Direct Marketing* 11(2): 7–19.
- 18 Pfeifer, P. and Robert, C. (2000) Modeling customer relationships as Markov chains. *Journal of Interactive Marketing* 14(2): 43–55.
- 19 Curry, J. and Curry, A. (2000) *The Customer Marketing Method*. New York: Free Press.
- 20 Zeithaml, V., Rust, R.T. and Lemon, K.N. (2001) The customer pyramid: Creating and serving profitable customers. *California Management Review* 43(4): 118–142.
- 21 Hwang, H., Jung, T. and Suh, E. (2004) An LTV model and customer segmentation based on

- customer value: A case study on the wireless telecommunication industry. *Expert Systems with Applications* 26(2): 181–188.
- 22 Myers, J.H. and Taubers, E. (1977) *Market Structure Analysis*. Chicago, IL: AMA.
- 23 Harald, H., Werner, F. and Markus, P. (2004) Market segmentation by maximum likelihood clustering using choice elasticities. *European Journal of Operational Research* 154(3): 779–786.
- 24 Kuo, R.J., Chang, K. and Chien, S.Y. (2004) Integration of self organizing feature maps and genetic algorithm based clustering method for market segmentation. *Journal of Organizational Computing and Electronic Commerce* 14(1): 43–60.
- 25 Balakrishnan, P.V., Cooper, C.M., Jacob, V.S. and Lewis, P.A. (1996) Comparative performance of the FSCL neural net and *K*-means algorithm for market segmentation. *European Journal of Operational Research* 93(2): 346–357.
- 26 Hruschka, H. and Natter, M. (1999) Comparing performance of feedforward neural nets and *K*-means for cluster-based market segmentation. *European Journal of Operational Research* 114(2): 346–353.
- 27 Green, P.E. and Tull, D.S. (1978) *Research for Marketing Decisions*, 4th edn. Englewood Cliffs, NJ: Prentice-Hall.
- 28 Galguera, L., Luna, D. and Mendez, M.P. (2006) Predictive segmentation in action Using CHAID to segment loyalty card holders. *International Journal of Market Research* 48(4): 459–479.
- 29 McCarty, J.A. and Hastak, M. (2007) Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research* 60(6): 656–662.
- 30 Boone, D.S. and Roehm, M. (2002) Retail segmentation using artificial neural networks. *International Journal of Research in Marketing* 19(3): 287–301.
- 31 Hruschka, E.R., Campello, R.J.G.B., Freitas, A.A. and de Carvalho, A.C.P.L.F. (2009) A survey of evolutionary algorithms for clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 39(2): 133–155.
- 32 Kuncheva, L.I. and Bezdek, J.C. (1997) Selection of cluster prototypes from data by a genetic algorithm. In: *Proceedings of 5th European Congress on Intelligent Techniques and Soft Computing*, <http://pages.bangor.ac.uk/~mas00a/papers/lkjbEUFIT97.pdf>, accessed 31 August 2011, pp. 1683–1688.
- 33 Estivill Castro, V. and Murray, A.T. (1997) Spatial clustering for data mining with genetic algorithms. In: *Proceedings of International ICSC Symposium on Engineering of Intelligent Systems*, http://neuro.bstu.by/ai/To-dom/My_research/Papers-0/For-research/N-networks/MiniMaxNN/estivill.pdf, accessed 31 August 2011, pp. 317–323.
- 34 Murty, C.A. and Chowdhury, N. (1996) In search of optimal clusters using genetic algorithms. *Pattern Recognition Letters* 17(8): 825–832.
- 35 Krishna, K. and Murty, N. (1999) Genetic *K*-means algorithm. *IEEE Trans on Systems, Man and Cybernetics – Pt. B* 29(3): 433–439.
- 36 Krovi, R. (1992) Genetic algorithm for clustering: A preliminary investigation. In: *Proceedings of the 25th Hawaii International Conference on System Sciences*, Vol. 4, http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=183445&tag=1, accessed 31 August 2011, pp. 540–544.
- 37 Lucasius, C.B., Dane, A.D. and Kateman, G. (1993) On *k*-medoid clustering of large data sets with the aid of a genetic algorithm: Background, feasibility and comparison. *Analytica Chimica Acta* 282(3): 647–669.
- 38 Bandyopadhyay, S. and Maulik, U. (2001) Nonparametric genetic clustering: Comparison of validity indices. *IEEE Trans on Systems, Man, and Cybernetics – Pt. C* 31(1): 120–125.
- 39 Ma, P.C.H., Chan, K.C.C., Yao, X. and Chiu, D.K.Y. (2006) An evolutionary clustering algorithm for gene expression microarray data analysis. *IEEE Transactions on Evolutionary Computation* 10(3): 296–314.
- 40 Bandopadhyay, S. and Maulik, U. (2002) An evolutionary techniques based on *k*-means algorithm for optimal clustering in R^n . *Information Sciences* 146(1–4): 221–237.
- 41 Cowgill, M.C., Harvey, R.J. and Watson, L.T. (1999) A genetic algorithm approach to cluster analysis. *Computational Mathematics and Its Applications* 37(7): 99–108.
- 42 Casillas, A., González de Lena, M.T. and Martínez, R. (2003) Document clustering into an unknown number of clusters using a genetic algorithm. In: V. Matoulek and P. Mautners (eds.) *Proceedings of the International Conference on Text Speech and Dialogue, LNCS*. Vol. 2807, Berlin Heidelberg: Springer Verlag, pp. 43–49.
- 43 Cole, R.M. (1998) Clustering with genetic algorithms. MSc Thesis, University of Western Australia, Australia.
- 44 Hruschka, E.R. and Ebecken, N.F.F. (2003) A genetic algorithm for cluster analysis. *Intelligent Data Analysis* 7(1): 15–25.
- 45 Tseng, L.Y. and Yang, S.B. (2001) A genetic approach to the automatic clustering problem. *Pattern Recognition* 34(2): 415–424.
- 46 Kaufman, L. and Rousseeuw, P.J. (1990) *Finding Groups in Data – An Introduction to Cluster Analysis*, Wiley Series in Probability and Mathematical Statistics, New York: Wiley.
- 47 Davies, D.L. and Bouldin, D.W. (1979) A cluster separation measure. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 1(2): 224–227.
- 48 Calinska, R.B. and Harabasz, J. (1974) A dendrite method for cluster analysis. *Communications in Statistics* 3(4): 1–27.
- 49 Fader, S., Hardie, B. and Lee, K.L. (2005) RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research* 42(4): 415–430.
- 50 Aeron, H., Kumar, A. and Janakiraman, M. (2010) Application of data mining techniques for customer lifetime value parameters: A review. *International Journal of Business Information Systems* 6(4): 514–529.

APPENDIX

The total columns included in the data analysis and their description are given below:

1. ACCNTNMB: Donor ID
2. CONTRFST: First contribution amount
3. CONLARG: Largest contribution amount
4. Totmonthlrgstcont: Time measured from 6/95 when largest contribution was made
5. CNTRLIF: Amount contributed lifetime
6. SLTMLIF: Number of time solicited lifetime
7. CNDOL1: Latest contribution
8. CNDOL2: 2nd latest contribution
9. CNDOL3: 3rd latest contribution
10. CNDOL4: 4th latest contribution
11. CNDOL5: 5th latest contribution
12. CNDOL6: 6th latest contribution
13. CNDOL7: 7th latest contribution
14. CNDOL8: 8th latest contribution
15. CNDOL9: 9th latest contribution
16. CNDOL10: 10th latest contribution
17. totmonthcont1: Latest contribution time measured from June 1995
18. totmonthcont2: 2nd latest contribution time measured from June 1995
19. totmonthcont3: 3rd latest contribution time measured from June 1995
20. totmonthcont4: 4th latest contribution time measured from June 1995
21. totmonthcont5: 5th latest contribution time measured from June 1995
22. totmonthcont6: 6th latest contribution time measured from June 1995
23. totmonthcont7: 7th latest contribution time measured from June 1995
24. totmonthcont8: 8th latest contribution time measured from June 1995
25. totmonthcont9: 9th latest contribution time measured from June 1995
26. totmonthcont10: 10th latest contribution time measured from June 1995
27. totmonthfirstcont: Time of first contribution measured from June 1995
28. CNTMLIF: Number of times contributed lifetime
29. last_contr: Time between first contribution and last contribution