Paper 130-2012

Comparison of K-means, Normal Mixtures and Probabilistic-D Clustering for B2B Segmentation using Customers' Perceptions

Satish Garla, Goutam Chakraborty, Oklahoma State University, Stillwater, OK, US
Gary Gaeth, University of Iowa, Iowa City, Iowa, US

ABSTRACT

Cluster Analysis is a popular technique used by businesses and analysts for market segmentation. For segmentation, clustering is used to split customers in a market into meaningful groups such that the customers within a group are similar and customers between the groups are dissimilar. Several clustering methods and numerous clustering algorithms are available in existing software packages and new ones frequently appear in the literature. These methods and algorithms vary depending on how the similarity between observations is defined or on other assumptions about shapes of clusters, distributions of variables, etc. This paper describes a comparative study of three clustering methods (K-means, Normal Mixtures and Probabilistic-D) for segmenting business-to-business (B2B) customers using their perceptions.

The "hard" clustering techniques such as K-means follow a deterministic approach in calculating cluster membership whereas the "soft" clustering techniques like Normal Mixtures calculate a degree of membership or probability for each customer to belong to a cluster. The Normal Mixtures technique, trained by the expectation-maximization algorithm, uses probability estimates via an iterative classification method. A new SAS® macro was developed for application of probabilistic-D technique. The macro calculates probability of cluster membership using the Euclidean distance of each observation from cluster centers found by k-means. These two soft clustering techniques are compared with the much widely used K-means technique. The results from each method are evaluated based on purity and cluster profiles. SAS® Enterprise Miner is used for K-Means and Probabilistic-D clustering and for profiling clusters while JMP® Pro 9 is used for Normal Mixtures. Our results show that a better understanding of markets can be achieved using soft clustering techniques.

INTRODUCTION

Cluster analysis is commonly used by market researchers as a classification tool to segment customer markets. Clustering creates groups of persons, products or events which can be used to determine managerial strategy, or are commonly the target of further analysis [15]. B2B markets are characterized by many factors such as complex buyer decision making, complex product attributes and trade-offs among such attributes as well as joint decision making for purchasing decisions. These characteristics can make understanding B2B customers a more demanding task than business-to-consumer customers [13]. In these markets, suppliers must carefully consider the nature and characteristics of their customers in order to satisfy them [8].

Several clustering methods and numerous clustering algorithms have been developed by statisticians, and are available in the literature. These methods and algorithms vary depending on how similarity between observations is defined as well as on other assumptions about shapes of clusters, distributions of variables, etc. At a high level, clustering techniques can be divided into two groups: classical (hard or deterministic) cluster analysis and probabilistic (fuzzy or soft) cluster analysis [2]. A number of researchers performed comparison studies evaluating the advantages and disadvantages of each approach as applied to data from various fields. In this paper we provide a comparative study of three clustering methods. Segments obtained by using hard clustering technique (k-means) are compared against the segments obtained from soft clustering techniques (probabilistic-D and Normal Mixtures). Our objective is not pick a winner based on this limited comparison using a single data set. Rather, we investigate whether there is additional value to researchers and businesses, by applications of these three techniques in a B2B domain.

K-MEANS

K-means algorithm is one of the most widely used hard clustering techniques. This is an iterative method where you specify the number of clusters beforehand. In this approach, each observation has 100% chance of belonging to one and only one cluster [12]. However, during the iterative process, an observation can shift from one cluster to another. The algorithm works as follows:

- ✓ Specify the number of clusters (k in k-means)
- ✓ Randomly select k cluster centers in the data space
- ✓ Assign data points to clusters based on the shortest Euclidean distance to the cluster centers

- ✓ Re-compute new cluster centers by averaging the observations assigned to a cluster
- ✓ Repeat above two steps until convergence criterion is satisfied

In general, this technique produces exactly k-clusters that are distinct to the greatest possible extent [7]. The advantage of this method is its capability to handle large data sets and can work with compact clusters. The major limitation of this technique is the requirement to specify the number of clusters beforehand and its assumption that clusters are spherical. This method is also sensitive to outliers and noise in the data [2]. SAS Enterprise Miner automatically selects the number of clusters (k starting points) by first running a hierarchical clustering on a sample of data using Cubic Clustering Criterion (CCC) and then uses the results form that step as an input to run k-means method [5].

PROBABILISTIC-D

Probabilistic-D clustering is an iterative soft clustering technique in which the cluster memberships of a data point are based on the distances (typically Euclidean) from the cluster centers. According to Israel and Iyigun (2008: p.5), in probabilistic-D (distance) clustering, "Given clusters, their centers and the distances of data points from these centers, the probability of cluster membership at any point is assumed inversely proportional to the distance from (the center of) the cluster in question." In this iterative approach, the cluster centers are updated as convex combinations of data points and this continues until the centers stop changing. This approach is similar to k-means, however, in this case the cluster assignment is "soft" and probabilities of cluster membership for each data point (i.e., consumer) are calculated [3]. This method is considered to be robust, insensitive to outliers and works best when cluster sizes are about equal [2].

In SAS Global Forum 2011, a SAS macro was published that discussed a way of implementing probabilistic-D clustering technique in SAS Enterprise Miner® [5]. This macro utilizes the distances calculated by the k-means algorithm to calculate cluster membership probabilities. We used this macro in order to segment the customers using Probabilistic-D clustering technique.

NORMAL MIXTURES

In JMP Normal Mixtures technique is trained by the popular Expectation Maximization (EM) algorithm. This technique is an iterative optimization method that estimates probabilities for each observation to belong to each cluster [4]. In situations where clusters overlap, assigning an observation to one cluster is debatable. Normal mixtures technique is thought to be especially useful in such situations.

The EM algorithm consists of two steps: Expectation (E) step and the Maximization (M) step. In the Expectation(E) step input partitions are selected similar to the k-means technique. In this step each observation is given a weight or expectation for each partition. In the second step, Maximization (M), the initial partition values are changed to the weighted average of the assigned observations, where the weights are those identified from E-step. This cycle is repeated until the partition values do not change significantly [13]. EM assumes that the joint probability distribution of the clustering variables can be approximated by a mixture of multivariate Normal distributions, which represent different clusters [12]. The EM algorithm is a very efficient and robust procedure for learning parameters from observations. This algorithm is also considered to be powerful in computing maximum likelihood estimates with incomplete data [17].

DATA

In this study we used data collected from a survey conducted by a supplier of hydraulic and pneumatic products serving 50,000+ customers in the USA. We cannot disclose the name of the company for confidentiality reasons. The data collected from 1,005 customers capture customers' perceptions of important attributes in selecting a supplier for the hydraulic and pneumatic products. Table 1 shows the variables that were included in the mail survey and the measurement scale used.

Before using this data for analysis, the data was processed to eliminate outliers and records with missing values. This processing task is critical since the clustering techniques studied are sensitive to outliers and missing values. Hence we excluded the observations with missing values. We are finally left with 787 observations that can be used for segmentation. One of the variables from the survey, Overall Satisfaction Rating, was used for evaluating the purity of the segments. This variable was measured on an 11 point scale whereas the variables used for clustering were measured on a 9 point scale. It is not always possible to capture the true intentions of the responders from a survey due to the presence of bias in the data [1]. This bias could possibly be introduced due to response style behavior.

Response styles in questionnaire/survey data is defined as the systematic inclination of responders to answer questions based on some unknown effect other than the content of the question [15]. In SAS Global Forum 2011, a paper discussed the advantages of using double-standardization as a method to eliminate response style behavior [12]. We used the SAS macro developed by these authors for transforming the data before performing segmentation.

How important are the following issues to customers in choosing a supplier for hydraulic, pneumatic and related	Attribute	Scale	Not at all important	Extremely Important
The reliability of the supplier	reliab	9 point	1	9
The timeliness of the deliveries by the supplier	time	9 point	1	9
The availability of a large breadth of products to choose from	av_br	9 point	1	9
The availability of well documented technical specification	av_spec	9 point	1	9
The price of products	price	9 point	1	9
The credit policy of the supplier	credit	9 point	1	9
The availability of electronic payment/debit option	av_pay	9 point	1	9
The return policy of the supplier	return	9 point	1	9
The warranty coverage provided by the supplier	warranty	9 point	1	9
The ability to talk directly to a salespeople about your needs	talk_dir	9 point	1	9
Overall Satisfaction with the current supplier	satisf	11 point	0% satisfied	100% satisfied

Table 1 Important factors for customers in selecting a supplier for the hydraulic and pneumatic products

RESULTS

K-MEANS

SAS Enterprise Miner was used for performing K-means analysis. Hierarchical clustering (Ward method) was used for identifying the number of clusters to input to K-means technique. The Ward method identified a 4 cluster solution. K-Means analysis from SAS generates four clusters as shown in figure 1. The results show a fairly equal distribution of customers in each of the four segments. For profiling clusters we used SAS Enterprise Miner ® 6.2. Using double-standardized variables for profiling clusters would likely reveal hard to interpret cluster definitions. Therefore, we used the raw variables for profiling each cluster.

The Segment profile node in SAS Enterprise Miner identifies important variables for each segment by building a decision tree to predict segment membership using the clustering variables as input. Figure 2 shows the worth (importance) of the variables for the K-means clusters. The panels within the figure are arranged from largest (segment value =3) to smallest (segment value =2).

The availability of electronic payment and reliability has the highest worth (most important in the decision tree model) to predict segment 3 members. For segment 1, credit policy, price and return policy of the supplier are three most important predictive variables. Credit policy and return policy of the supplier emerge as the two most important predictive variables for segment 4. For segment 1 availability of detailed technical specification followed by availability of large breadth of products seem to be the two important predictive variables.

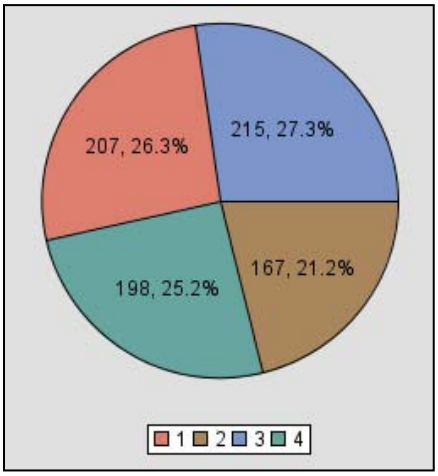

Figure 1 Cluster Sizes - (K-Means)

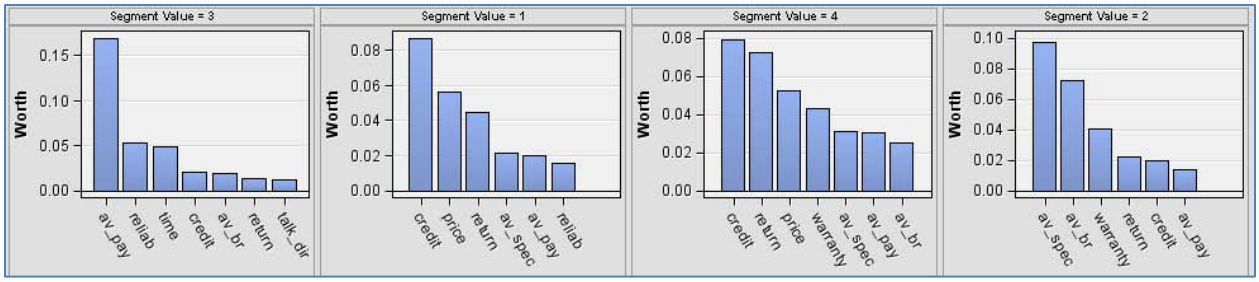

Figure 2 Important factors in the clusters generated using K-Means

PROBABILISTIC-D

Probabilistic-D clustering technique is not currently available in any commercial package. We used a SAS® macro developed by SAS users whose work was presented at SAS Global Forum 2011 [5]. In this approach the K-means algorithm is first run in SAS Enterprise Miner®. The score code generated from the k-means technique is used in the macro to calculate cluster membership probabilities. Figure 3 shows the process diagram from SAS Enterprise Miner ®6.2.

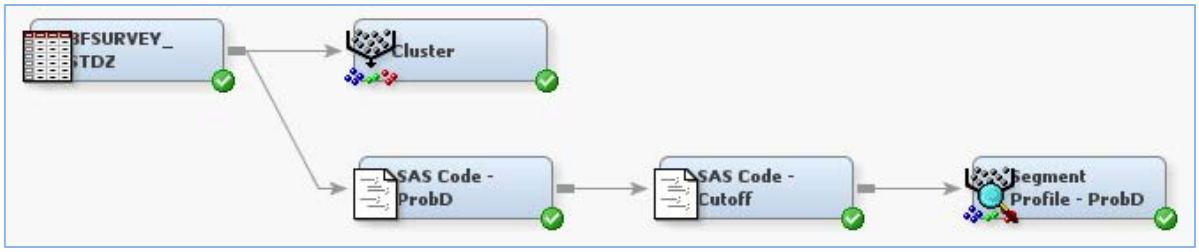

Figure 3 Process Flow for Probabilistic-D Clustering in SAS Enterprise Miner ®

The macro can calculate probabilities using either Euclidean distance or exponential distance. For this study we used only the Euclidean distance measure.

Though we have probabilities for cluster membership, in order to compare Probabilistic-D clustering results with other cluster techniques each observation needs to be assigned to only one cluster. We explored various cut-off probabilities for determining unique cluster membership for each observation. After some trial-and-error, we identified a reasonable classification with 601 observations at a probability cut-off of 0.33 that resulted in roughly equal distribution of observations into the four clusters as shown in Figure 4. Figure 5 shows important variables for the clusters identified using this technique.

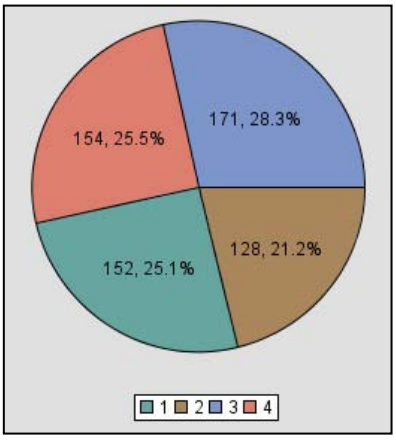

Figure 4 Cluster Sizes (Probabilistic-D)

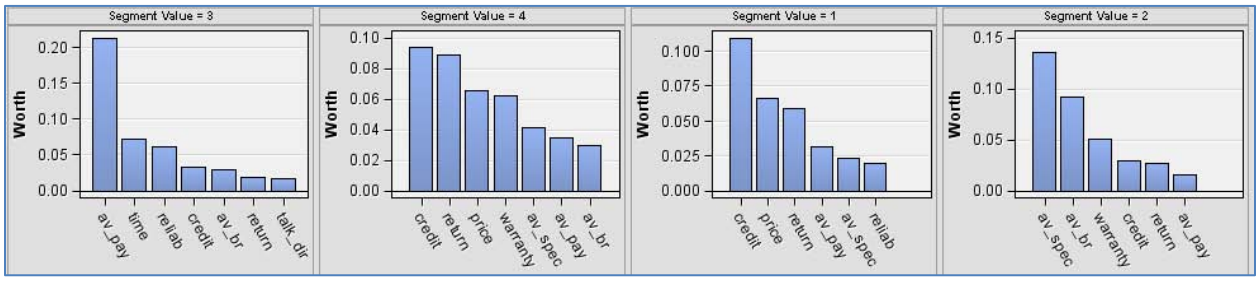

Figure 5 Important factors in the clusters generated using Probabilistic-D clustering

For segment 3, availability of electronic payment, timeliness of deliveries and reliability of the supplier emerge to be most important factors. For segment 4, many attributes including credit, return, price, warranty policy turn out to be important. For segment 1, credit policy, price and return policy of the supplier emerge as important. For segment 2, availability of detailed specifications, availability of large breadth of brads turn out to be important

NORMAL MIXTURES

JMP Pro ® 9.0 was used for performing Normal Mixture technique. As explained earlier this method also starts with a predefined value for the number of clusters. We used four as the number of clusters as identified using Ward method. Figure 6 shows the number of clusters identified by Normal Mixtures method. Even in this method the results show a fairly equal distribution of customers in each of the four segments. The two steps in the Expectation Maximization algorithm can be controlled using the Tours and Maximum Iterations options available in the JMP software. A value for Tours helps in testing different number of independent restarts of estimation process.

A value for the maximum number of iterations property controls that number of iterations in the second stage that are used for convergence. These two properties significantly impact the cluster sizes. The profiles of the clusters as identified using Segment profile node in SAS Enterprise Miner are shown in Figure 7.

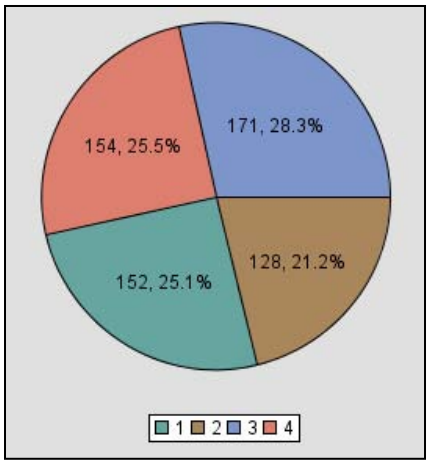

Figure 6 Cluster Sizes (Normal Mixtures)

For predicting the largest cluster (segment value =3), the important variables are ability to talk directly to sales representatives, followed by timeliness and reliability of the supplier. This variable was not identified as the most important factor by either K-means or probabilistic-d clustering. The same factor is also identified as dominant in the second largest segment (segment value=1). For predicting segment 2, timeliness of deliveries and reliability emerge as two most important variables. For Segment 4, important predictive variables include return policy, credit policy and availability of detailed specifications.

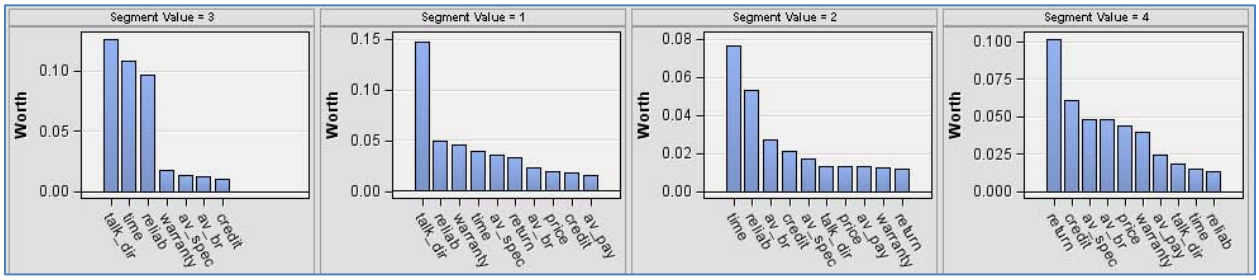

Figure 7 Important factors in the clusters generated using Normal Mixtures

COMPARISON

The patterns of the worths (importance for predicting each segment) from the segment profile paint quite a diverse picture for the four segments obtained by the three methods. To provide readers with deeper insights into these segments, we have also reported the traditional profiles (means of variables) for each segment for each method as shown below. The largest and the smallest mean ratings for each attribute are colored in red and blue respectively.

Segment	reliab	time	av_br	av_spec	price	credit	av_pay	return	warranty	talk_dir
1	8.76	8.65	6.97	8.20	6.86	4.21	2.45	5.83	7.64	8.55
2	8.63	8.65	5.35	6.23	7.81	5.44	2.59	6.01	6.77	8.14
3	7.98	8.01	7.17	7.62	7.63	6.57	5.61	7.25	7.77	8.03
4	8.67	8.64	7.53	8.34	8.52	7.48	2.21	8.05	8.52	8.40

Table 2 Mean rating for all the variables across the clusters identified by k-means

Segment	reliab	time	av_br	av_spec	price	credit	av_pay	return	warranty	talk_dir
1	8.83	8.74	7.10	8.30	6.76	4.03	2.21	5.72	7.66	8.64
2	8.67	8.76	5.41	6.02	7.82	5.28	2.60	6.01	6.68	8.19
3	8.03	7.99	7.40	7.75	7.69	6.91	6.08	7.39	7.81	8.07
4	8.72	8.71	7.72	8.47	8.65	7.81	2.25	8.25	8.67	8.60

Table 3 Mean rating for all the variables across the clusters identified by Probabilistic-D

Segment	reliab	time	av_br	av_spec	price	credit	av_pay	return	warranty	talk_dir
1	8.20	8.20	6.39	7.17	7.41	5.73	3.41	6.38	7.12	7.42
2	8.04	7.96	6.34	7.24	7.40	5.29	2.71	6.39	7.45	8.09
3	9.00	9.00	7.05	7.97	7.71	5.90	3.21	6.74	8.00	9.00
4	8.97	8.97	8.34	8.97	8.97	8.00	4.48	8.97	8.97	8.97

Table 4 Mean rating for all the variables across the clusters identified by Normal Mixtures

From Table 2 we can see that the highest and the lowest mean ratings for each variable are spread out across the clusters. From Table 3 we see that Probabilistic-D technique also has a similar kind of distribution as observed in the case of K-Means. In the case of Normal Mixtures, we can see from Table 4 that cluster 4 has the highest mean ratings for most of the variables and cluster 2 has the lowest mean for most of the variables.

In addition to comparing mean values, difference in the results from the clustering methods can also be identified by looking at the range of means (maximum mean rating – minimum mean rating) for each attribute across the clusters. Table 5 shows the range of the attribute means reported in Table 2, 3 and 4. For each attribute, the largest range value is highlighted in blue color.

Method	reliab	time	av_br	av_spec	price	credit	av_pay	return	warranty	talk_dir
K-means	0.78	0.64	2.18	2.12	1.67	3.27	3.41	2.22	1.75	0.51
Probabilistic-D	0.80	0.76	2.31	2.44	1.89	3.78	3.87	2.52	1.99	0.57
Normal Mixtures	0.96	1.04	1.99	1.80	1.57	2.71	1.77	2.59	1.85	1.58

Table 5 Range of means for each variable from in all the three techniques

Looking at the means and the ranges of the means for each attribute, we can observe that the probabilistic-D and Normal Mixtures tend to separate the means across clusters better than the k-means. Better separation makes profile of segments easier to understand and easier to act upon for developing tailored marketing communications. We have also seen differences in terms of the important predictive variables from the segment profile nodes. Which of these segments are more meaningful and actionable can truly be evaluated by experts in the B2B domain. However, we have used one of the variables in the survey to evaluate the purity of these clusters. The survey included a question about overall satisfaction with the current supplier, with an 11-point measurement scale (0-0%satisfied, 11 –100% Satisfied). We combined the top two response boxes (10, 11) and measured the percentage of customers in each cluster who are considered highly satisfied with the current supplier. Table 6 shows these percentages.

Cluster	K-means	Probabilistic-D	Normal Mixtures
1	40.58%	42.11%	36.73%
2	36.53%	35.94%	34.68%
3	40.00%	42.11%	43.57%
4	42.42%	43.51%	51.02%

Table 6 Percent classification of customers who rated high on overall satisfaction measure

Normal Mixtures seems to separate the clusters better than the separation obtained from other methods. The K-means method classifies the satisfied customers almost equally across the clusters. The results from Probabilistic-D clustering are similar to the results obtained from k-means method.

CONCLUSION

This study presents a comparison of different clustering methods in the B2B domain. Our results show wide differences in the profiles of clusters generated from each method. In most practical applications, the shapes of clusters, the distributions of clustering variables, number of clusters, whether clusters overlap or not, etc., are unknown. Therefore, it is not possible to theoretically justify one clustering method over another because of the assumptions of each of the clustering methods. Thus, at the end of the day, the utility of each clustering method has to be evaluated by domain experts via judging the usefulness of each cluster solution based on profiles and via field studies to test the marketing effectiveness of each cluster solutions using control and test groups. Using of a descriptor variable, such as the overall satisfaction that was not used in deriving the clusters, can also help to a certain extent in validating the cluster results. Given this criterion, our analyses show that Normal Mixtures is performing slightly better when compared to other methods. This suggests that analysts may gain deeper insights by including Normal Mixtures along with other clustering techniques.

REFERENCES

[1]Bachman, J.G., & O'Malley, P.M. (1984). *Yea-saying, nay-saying, and going to extremes: Blackwhite differences in response styles.* Public Opinion Quarterly, 48, 491-509.

[2]Budayan, C. (2008). *Strategic group analysis: Strategic perspective, differentiation and performance in construction*. Doctoral dissertation, Middle East Technical University

[3]Budayan,C.,Dikemen, I., & Birgonul, T. (2009). Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping. Expert Systems with Applications. www.elsevier.com/locate/eswa

[4]Dellaert, F. (2002). The Expectation Maximization Algorithm. College of Computing, Georgia Institute of Technology.

[5]Dey., Garla, Chakraborty, (2011). *Comparison of Probabilistic-D and k-Means Clustering in Segment Profiles for B2B Markets*. SAS Global Forum

[6]Girish, P., & David, W.S., (1983). *Cluster Analysis in Marketing Research: Review and Suggestions for Application*. American Marketing Association. http://www.jstor.org/stable/3151680

[7]Granzow, M., Berrar, D., Dubitzky, W., Schuster, A., Azuaje, F.J., & Eils, R. (2001). *Tumor Classification by Gene Expression Profiling: Comparison and Validation of Five Clustering Methods.* ACM-SIGBIO Newsletters.

[8]Hosseini, S., Maleki, A. & Gholamian, M. (2010). *Cluster analysis using data mining approach to develop CRM methodology to access the customer loyalty*. Expert Systems with Applications, 37, pp 5259 – 5264.

[9]Israel, A., & Iyigun, C. (2008). *Probabilistic D Clustering*. Journal of Classification, 25, pp 5 – 26

[10]Iyigun, C., & Israel, A. (2010). *Semi-supervised probabilistic distance clustering and the uncertainty of classification* in A Fink et al., *Advances in Data Analysis, Data Handling and Business Intelligence, Studies in Classification, Data Analysis, and Knowledge Organization*. Berlin: Springer.

[11]McLachlan, G. J. & Krishnan, T. (1998). *The EM Algorithm and Extensions*. John Wiley and Sons, Hoboken, New Jersey.

[12]Pagolu, Chakraborty, (2011). *Eliminating Response Style Segments in Survey Data via Double Standardization Before Clustering*. SAS Global Forum

[13]Nathiya, G., Punitha, .S.C., & Punithavalli, M. (2010). *An Analytical Study on Behavior of Clusters Using K Means, EM and K* Means Algorithm*. International Journal of Computer Science and Information Security, Vol.7, No.3.

[14]Paul, H., & Matthew, H., *White Paper: Market Segmentation in B2B Markets*. Online Publication: B2BInternational.com (http://www.b2binternational.com/publications/white-papers/b2b-segmentation-research/)

[15]Paulhus, D. L. (1991). *Measurement and control of response bias*. In J.P. Robinson, P.R. Shaver, & L.S. Wrightman (Eds.), *Measures of Personality and Social Psychological Attitudes* (Vol. 1). San Diego, CA: Academic Press.

[16]Simkin, L. (2008). *Achieving market segmentation from B2B sectorisation*. Journal of Business & Industrial Marketing, 23(7), pp 464 – 474

[17]Flury,B.& Zoppe,A. (2000). *Exercises in EM*. The American Statistician. Aug 2000.

TRADEMARKS

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Satish Garla, Oklahoma State University, Stillwater OK, Email: satish.garla@okstate.edu

Satish Garla is a Master's student in Management Information Systems at Oklahoma State University. He has three years of professional experience as Oracle CRM Consultant. He is SAS Certified Advanced Programmer for SAS® 9 and Certified Predictive Modeler Using SAS® Enterprise Miner 6.1

Dr. Goutam Chakraborty, Oklahoma State University, Stillwater OK, Email:goutam.chakraborty@okstate.edu

Goutam Chakraborty is a professor of marketing and founder of SAS and OSU data mining certificate program at Oklahoma State University. He has published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He is also a Business Knowledge Series instructor for SAS.

Dr. Gary Geath, University of Iowa, Iowa City, Iowa
Email: gary-gaeth@uiowa.edu

Gary Gaeth is Professor of Marketing and holds the Cedar Rapids Chair of Business. He previously served as Associate Dean of the School of Management for 12 years, where he and his leadership team were responsible for the 4 different MBA programs, including a Full-time Program of nearly 200 students, a MBA for Professionals of nearly 800 students and Executive MBA programs in Des Moines, Hong Kong, Beijing and most recently Italy. His theory-based research focuses on mathematical models of consumer decision making and his applied research on strategic management in the service businesses. Gary has published widely in marketing, retailing and decision making journals and has consulted extensively with companies in several diverse industries. He has received the Collegiate Teaching Award, as well as Teacher of the Year award from Executive MBA's, Full-Time MBA's, and Ph.D. students.