Paper 193-2011

Comparison of Probabilistic-D and k-Means Clustering in Segment Profiles for B2B Markets

Dipanjan Dey, Satish Garla, Goutam Chakraborty, Oklahoma State University, Spears School of Business, Stillwater, OK, U.S.

ABSTRACT

Customer segmentation is a critical business analysis tool that allows organizations to build customer profiles and plan marketing efforts to satisfy the varying demands of different segments. The objective of the present study is to empirically explore the concept of probabilistic-D clustering for segment profiling in a business-to-business (B2B) market. A SAS® macro that can be used on any data set for application of this technique is developed and reported. To the best of our knowledge, probabilistic-D technique has never been empirically tested in a business environment. It was compared with the widely used k- means clustering technique. Findings indicate a better explanation of customer segment profiles using probabilistic-D clustering because k-means seems to force a high percent of observations in segments where cluster membership probabilities are less than 50% as calculated by probabilistic-D clustering macro. These observations are likely to be non-responsive (or less responsive) to the marketing efforts directed towards respective k-means segments. Using a probabilistic-D clustering approach these observations can be targeted differently to improve the effectiveness of marketing communication and promotions.

INTRODUCTION

Most markets as well as customers are heterogeneous in their needs and preferences (Clarke, 2009). In industrial markets, suppliers must carefully consider the nature and characteristics of their customers in order to satisfy them (Hosseini, Maleki & Gholamian, 2010). Segmentation as a technique for forming customer groups for effective targeting is a widely researched area in marketing (Simkin, 2008). Cluster analysis is a popular tool to segment markets. Simply stated, it is a technique for separation of customers into different groups such that each group of customers is collectively different from the customers in the other groups. Many methods of cluster analysis are available in the literature. But on a broad basis, clustering techniques can be divided into two groups classical (hard or deterministic) cluster analysis and probabilistic (fuzzy or soft) cluster analysis (Budayan, 2008). A number of studies carried out in different fields compare the performance of these two different clustering approaches (Budayan, Dikmen, & Birgonul, 2008). In a majority of these comparison studies, fuzzy clustering is discussed as the most popular form that has been adopted in diverse fields, presumably because it adds valuable diagnostics over hard clustering (Ozer, 2001). The purpose of this paper is to introduce a relatively unexplored field of soft clustering technique for market segmentation. This technique is known as probabilistic-D clustering (Israel and Iyigun, 2008). We attempt to empirically test whether this new approach adds any value to the existing literature on market segmentation by making a comparison of segment profiles obtained from commonly used hard clustering (k-means) using SAS Enterprise Miner 6.1 ® and probabilistic-D clustering using a new SAS® macro developed by the authors. Our objective is not to criticize any existing clustering techniques, but rather to investigate whether introduction of probabilistic-D clustering can explain the market complexities in a manner better suited to business needs. We demonstrate the utility of probabilistic-D clustering on survey-based data collected by a leading supplier of hydraulic and pneumatic products.

HARD CLUSTER ANALYSIS

The term “hard cluster” analysis refers to all clustering techniques where the assignment of observations to cluster is deterministic. Stated differently, in hard clustering techniques each observation has 100% chance of belonging to one and only one cluster. There are two main groups of clustering methods, hierarchical and non-hierarchical clustering, each with many different sub-methods and algorithms. In agglomerative hierarchical methods, each observation is initially assigned to its own cluster and then merged with others based on a similarity measure. The algorithm continues until all data points form a single cluster solution. While widely applied, hierarchical methods are less suitable for market segmentation because a priori there is no reason to expect the market segments to have a hierarchical structure. In non-hierarchical methods such as k-means, an iterative partitioning algorithm is used that does not impose a hierarchical structure (Budayan, 2008). We selected k-means, one of the most widely used clustering methods for segmentation, to compare with probabilistic-D clustering.

K-MEANS CLUSTER ANALYSIS

k-means cluster analysis is one of the most popular hard cluster analysis techniques (Blattberg et al., 2008). In a classic application of this technique, the number of clusters k must be pre-specified. The algorithm then selects cluster centers and each of the observations in the data is assigned to a particular cluster based upon the shortest Euclidean distance of the data point from the cluster centers. It is an iterative procedure; once observations are assigned to cluster centers, new cluster centers are created by averaging the observations assigned to a cluster. Distances from these new cluster centers are calculated for all observations, and the assignment of observations to clusters continues until a convergence criterion is satisfied (Budayan, 2008). This method has a number of advantages, such as its ability to handle large amounts of data points, and its ability to work with compact clusters (Budayan, 2008). However, it has its own set of limitations as well, such as the variables must be commensurable (Blattberg et al. 2008), the number of clusters should be known beforehand, and it is sensitive to outliers and noise (Budayan, 2008). In recent years, algorithms have been developed for an automatic (multi-stage) way of selecting the number of clusters, the k in k-means. For instance, in SAS Enterprise Miner 6.1 ® the number of clusters, k, is first determined by running a hierarchical clustering on a sample of data using CCC (cubic clustering criterion) and then running the k-means algorithm on the entire data set.

SOFT CLUSTER ANALYSIS

The term “soft cluster” analysis refers to all clustering techniques where assignment of observations to clusters is chance-based. In other words, in soft clustering techniques there is a chance that each observation could belong to any of the clusters. Thus, the probabilistic clustering technique assigns probabilities of cluster memberships to each observation; therefore, it is not deterministic. Soft clustering techniques overcome the limitation of forceful assignment of an observation to a single cluster and hence are more appealing in business situations where segments may not be clearly differentiable and may be overlapping in character (Chuang, Chiu, Lin, & Chen, 1999). Fuzzy C means clustering is the most commonly known type of soft clustering. However, we discuss here a relatively new and a simpler method of soft clustering, as described below.

PROBABILISTIC-D CLUSTER ANALYSIS

As per 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.”

If, P_k (x) = probability that the point x belongs to cluster C_k.

d_k (x) = distance of point x from cluster C_k

Then: p_k (x). d_k (x) = constant, depending on (x).

The clustering criterion being used here is Euclidean distances.

Mathematically as per Iyigun and Israel (2010):

$$P_k(\boldsymbol{x}) = \frac{\prod_{j \neq k} d_j(\boldsymbol{x})}{\sum_{i=1}^{K} \prod_{j \neq i} d_j(\boldsymbol{x})}$$

There can be many ways of operationalizing the distances. If exponential distance is considered, then as per Israel and Iyigun (2008):

P_k (x) e d_k (x) = constant, depending on (x)

Accordingly for calculation of probability in equation above, d_j (x) is replaced by e d_j (x).

Probabilistic-D clustering has all the advantages of generic soft clustering techniques over hard clustering techniques such as k-means. Fuzzy C Means (FCM) cluster analysis is the most well known and widely researched technique in soft clustering (Ozer, 2001). The main differences between FCM and probabilistic-D clustering is that while FCM determines the cluster centers as well as the distances between the cluster centers and observations simultaneously,

in Probabilistic-D clustering the cluster centers are determined first. Then, based on those cluster centers, the distances (Euclidean/Exponential) are calculated to assign probabilities of cluster membership. Our motivation to look for an approach other than FCM is as follows. First, FCM is known to be slow to converge, especially with large data sets (Chuang et al. 1999). Second, in spite of our best efforts, we could not find a macro or algorithm to readily apply FCM using SAS®. Israel and Iyigun (2008) argue that probabilistic-D clustering is a simpler process, is robust and gives a higher percentage of correct classifications. From a SAS® user point of view, application of probabilistic-D clustering should be easier because it can be built upon the familiar k-means output by extracting the distances from cluster centers and then using those distances to calculate the probabilities of cluster memberships. To this end, we developed and report a new SAS® macro that can be used easily with large data sets to calculate the probabilities of cluster memberships for a range of cluster solutions after executing k-means on SAS Enterprise Miner 6.1 ®.

DATA AND MEASUREMENT

A leading supplier of hydraulic and pneumatic products, located in the midwestern U.S., conducted a mail-survey of their customers. The survey contained a battery of questions to identify its customers' performance and satisfaction levels. The name of the organization and the variables directly related to the organization are suppressed to maintain client confidentiality. Sample size for the original survey consists of 1,068 data points. Ten variables were used for segmentation using cluster analysis. A nine point rating scale was used for variables used in segmentation. Appendix C contains the list of variables and the scales used. SAS Enterprise Miner 6.1 ® was used to perform analysis which primarily consists of k-means cluster analysis and probabilistic-D clustering using the newly developed SAS® macro. Using SAS Enterprise Miner 6.1 ®, preliminary data cleansing was done to filter out missing data and remove outliers. The final sample size was reduced to 911.

RESULTS

K-MEANS CLUSTERING

As explained earlier, SAS Enterprise Miner6.1 ® uses a two-step approach to automatically and rapidly decide *k* in k-means clustering using cluster features (Budayan, 2008). We used Ward's method in the first step of the two-step algorithm. Three relatively equal sized segments were obtained using the above procedure as shown in Figure 1. The means for each segmentation variable for each of the three clusters as well as for all 911 observations are shown in Table 1.

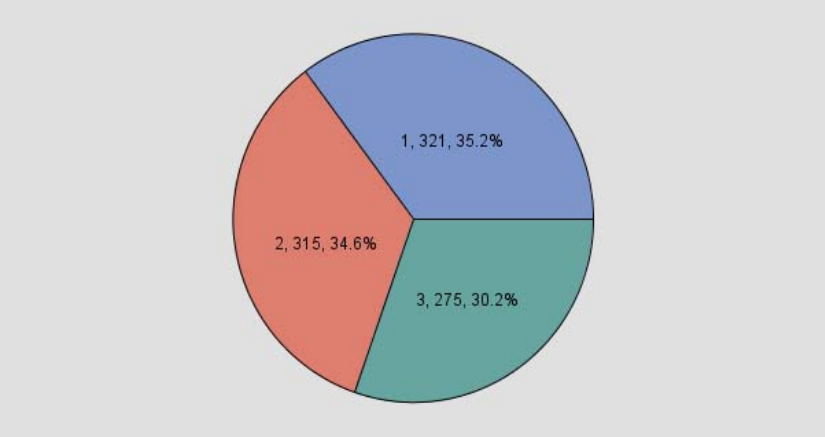


Figure 1. Segment Size

SEG	FREQ	av_br	av_pay	av_spec	credit	price	reliab	return	talk_dir	time	warranty
1	321	7.74	5.93	8.26	7.38	8.26	8.68	7.83	8.66	8.67	8.42
2	315	5.85	2.38	6.82	4.36	6.90	8.29	5.60	7.90	8.24	6.66
3	275	7.19	1.76	8.15	6.71	8.11	8.74	7.65	8.58	8.71	8.50
Total	911	6.92	3.44	7.73	6.14	7.74	8.56	7.00	8.37	8.53	7.84

Variable Description: Refer Appendix C

Table 1. Means of Variables Used in Segmentation

PROBABILISTIC-D CLUSTERING

On the same data set of 911 observations, probabilistic-D clustering was carried out with the help of the macro that is described in the Appendix A. In the macro, we used both the Euclidean and exponential distance for calculating probabilities of observations being in a particular segment (in our case, three segments). We note that in this data set, the exponential distance does not seem to add any value to the already existing k-means clusters. The probabilities assigned for observations to a particular segment via exponential distances were very high (either 1 or very close to 1). Based on these probabilities, the cluster membership assignments to the observations were essentially similar to the k-means results.

When we used Euclidean distance for calculating the probabilities, a very different scenario emerged. We arbitrarily decided that if the probability of any observation belonging to any segment is below 0.5, then we considered the observation a fuzzy case and did not include it for profiling that segment. We understand that the choice of a 50% probability cut-off is arbitrary. However, from a business perspective, a 50% cut-off seems to make sense. We found that 251 observations met that criterion (probability of membership to any cluster is less than 0.5) out of the total 911 observations. That left us with 660 observations that we believe can be unambiguously classified into one of the three clusters for profiling purposes. Figure 2 shows the frequency distribution of segment memberships of these 660 observations. The means for each segmentation variable for each of the three clusters as well as for all 660 observations are shown in Table 2.

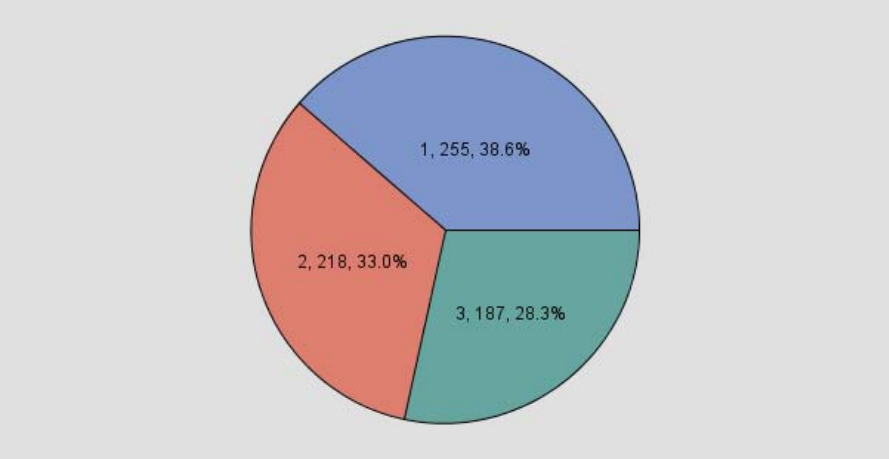


Figure 2. Segment Size

SEG	Freq	av_br	av_pay	av_spec	credit	Price	reliab	Return	talk_dir	time	Warranty
1	255	7.93	6.16	8.39	7.68	8.40	8.75	8.05	8.73	8.74	8.56
2	218	5.62	2.13	6.61	4.13	6.79	8.22	5.28	7.78	8.14	6.36
3	187	7.41	1.52	8.34	7.13	8.27	8.78	7.88	8.67	8.75	8.61
Total	660	7.02	3.52	7.79	6.35	7.83	8.58	7.09	8.40	8.54	7.85

Variable Description: Refer Appendix C

Table 2. Means of Variables Used in Segmentation

COMPARISON OF PROFILES FROM TWO METHODS

Table 1 and 2 describe the mean statistics of the variables for k-means and probabilistic-D clustering. Comparing the patterns of the means in the two tables, we can see a clear difference between them. The average values of cluster means seem to be higher and therefore clearer and stronger for Table 2. A similar conclusion can be graphically seen in terms of the variable worths reported in Figure 3 and Figure 4, which are outputs from the Segment Profile node in SAS Enterprise Miner following applications of k-means and probabilistic-D clustering. Thus, managerially it seems easier to interpret the profiles in Table 2 or Figure 4 and come up with appropriate marketing programs.

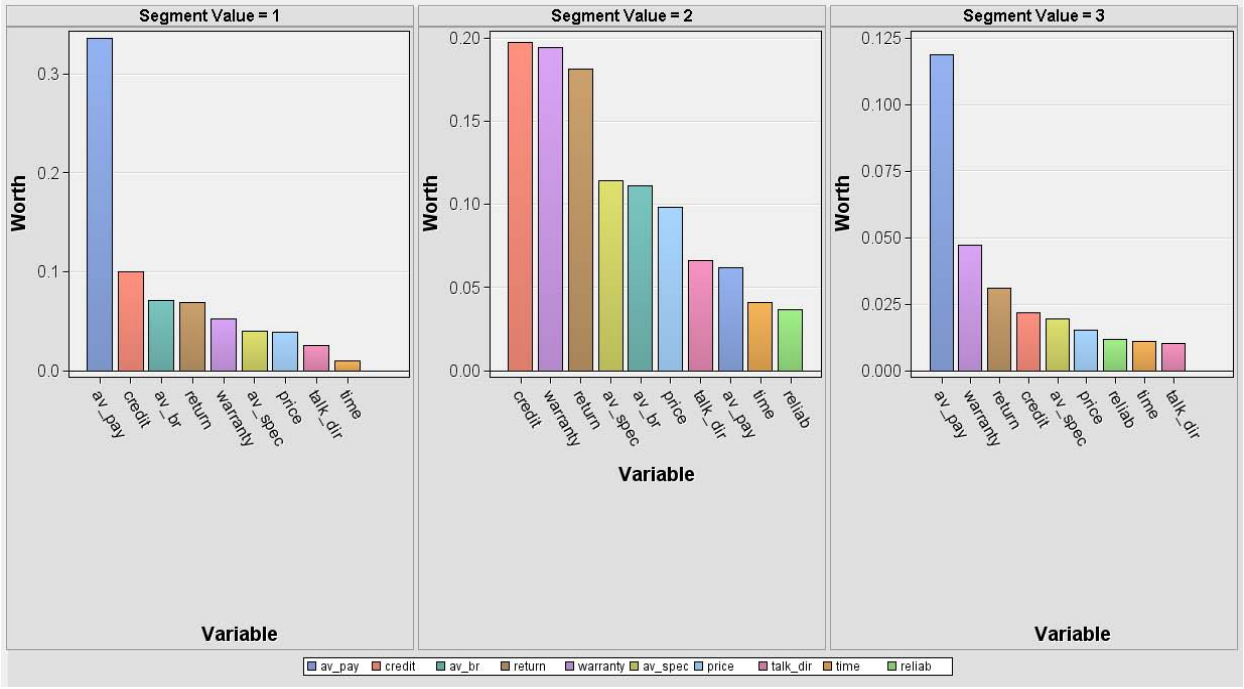


Figure 3 – Segment Profiles: K MEANS CLUSTER ANALYSIS (911 Observations)

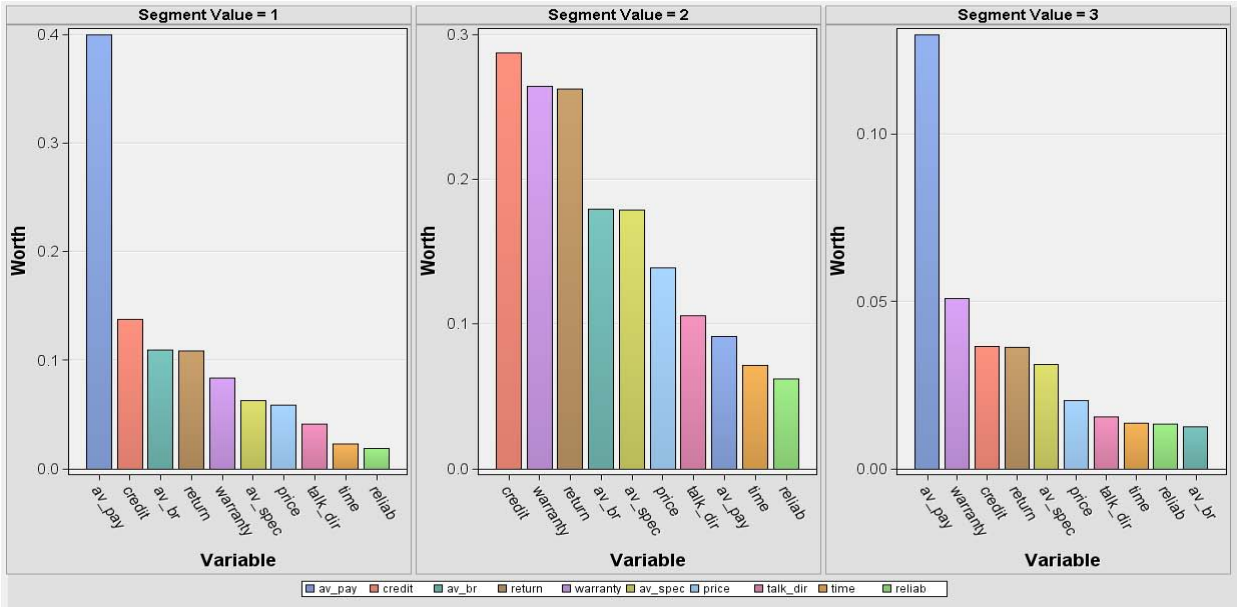


Figure 4 – Segment Profiles: Probabilistic-D Clustering – Euclidean Distance (660 Observations)

CONCLUSION

Our results show that the use of probabilistic-D clustering approach improved the interpretability and usability of segment profiles compared to those from k-means. The cutoff criterion of at least having a probability of 0.5 to be definitely assigned a cluster membership has several managerial implications. The higher the cut-off number, generally the more severe will be the reductions in number of observations that can be unambiguously classified into a cluster. On the other hand, the lower the cut-off number, the less clear are the differences in profiles between k-means and probabilistic-D clustering. Use of a 0.5 cut-off in this study seems to have produced a reasonable trade-off

between our desires to have maximum number of observations classified versus clearer segment profiles. In this study, more than one fourth of the sample was deemed to be fuzzy and not assigned to any cluster; the rest were assigned hard cluster membership. In order to unequivocally demonstrate whether our approach is better than assigning all observations to hard clusters, a field study is needed where customized communications using control and test groups from both methods can be tested for marketing effectiveness. While we do not have the luxury of conducting such a field study, we can argue, based on business sense that our approach is better as described below. If this client company used k-means, they would have assigned all 911 observations to one of the three clusters and sent targeted communications to all of them. However, the 251 fuzzy members in this data; may not really belong to any of the three clusters, and hence it would likely have been a waste of resources to target those via marketing communications.

As far as our future research directions are concerned, we only worked on the first part of Israel & Iyigun (2008) probabilistic-D clustering. They also suggested optimization of the distances for probability calculations and using joint distance function as a monitoring mechanism. As a next step our SAS® macro can be further developed to deal with the joint optimization using commonly available algorithms in SAS®.

APPENDIX

A. MACRO FOR CALCULATION OF PROBABILITIES USING EUCLIDEAN & EXPONENTIAL DISTANCE

Macro Name: PROBCLUSTER
Purpose: Macro to calculate the probability of a data point belonging to a cluster, i.e., the cluster membership is calculated given a data point, number of clusters and the distance measure of the point from each cluster center. The membership is calculated for both distances and exponential of distances.

How it Works: The macro takes as input the array of distances as calculated from any Clustering Algorithm like a K-Means Algorithm. The probability measure is calculated based on the simple principle: “the probability of membership is inversely proportional to the distance of the data point to the cluster center” (Israel & Iyigun, 2008: p.5).
Parameters: The macro has four parameters; two input and two output. Input parameters are the Distance Array and the Number of Clusters. The Output Parameters are two Arrays of Cluster Probabilities.
CARRAY= Array Name of the Cluster Distances
PARRAY_D= User Defined Array Name for Probabilities based on Distances
PARRAY_ED= User Defined Array Name for Probabilities based on Exponential distance values
SIZE= Number of Clusters
NOTE: If the distance from any of the cluster is zero, the macro calculates the probability as 1 for that cluster and 0 for all other clusters.

```
Options merror mprint mlogic;
%macro probcluster(CARRAY,PARRAY_D,PARRAY_ED,SIZE);
       array &PARRAY_D.&PARRAY_D.1-&PARRAY_D.&SIZE.;
       array &PARRAY_ED.&PARRAY_ED.1-&PARRAY_ED.&SIZE.;
       array D&CARRAY.[&SIZE.] _temporary_;

       length       DIST 8.; drop DIST;
       length       SCASE 8.; drop SCASE;
       length       i 8.;drop    i;
       length       j 8.;drop    j;

do i = 1 to &SIZE.;
       DIST=&CARRAY.[i];
       D&CARRAY.[i] = Exp(DIST);
       &PARRAY_D.[i]=0;
       &PARRAY_ED.[i]=0;
end;

/* Check if any of the distances are Equal to Zero (The datapoint is the
Cluster Center)*/
SCASE = 0;
```

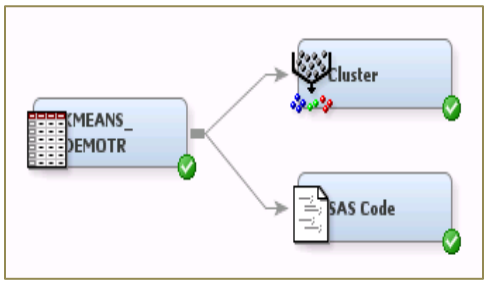
```
do i = 1 to &SIZE.;
   if &CARRAY.[i]=0 then SCASE = 1;
end;

/* If there are Zero Distances DO only the IF part otherwise DO the ELSE part*/
if SCASE >0
then
do;
       do i = 1 to &SIZE.;
              if &CARRAY.[i]=0 then
                     do;
              &PARRAY_D.[i] = 1;
                     &PARRAY_ED.[i] = 1;
                     end;
              else
                     do;
              &PARRAY_D.[i] = 0;
                     &PARRAY_ED.[i] = 0;
                     end;
       end;
end;
else
do;
       do i = 1 to &SIZE.;
              do j=1 to &SIZE.;
                     &PARRAY_D.[i]=&PARRAY_D.[i]+(&CARRAY.[i]/&CARRAY.[j]);
                     &PARRAY_ED.[i]=&PARRAY_ED.[i]+(D&CARRAY.[i]/D&CARRAY.[j]);
              end;
              &PARRAY_D.[i]=1/&PARRAY_D.[i];
              &PARRAY_ED.[i]=1/&PARRAY_ED.[i];
       end;
end;

%mend;
```

B. EXAMPLE

Below is the example code block to run PROBCLUSTER macro from SAS® Code Node in SAS Enterprise Miner®. A sample flow diagram is shown in Figure below.



The complete block shown below should be included in the SAS® Code Editor. Before including this code, make sure to complete running Cluster Node. The code below has two parts: macro definition and a DATA step. The call to the macro PROBCLUSTER is included in the DATA step. Paste the macro in the space identified as Placeholder1. Next, copy the entire Score Code generated in the Cluster Node and paste it in the space identified as Placeholder2. In the score code, identify the number of clusters by looking at **CLUSvads** array definition.

For example, for a three cluster solution the **CLUSvads** array definition looks like this,

```
arrayCLUSvads [3] _temporary_;
```

Enter the identified Size for the **SIZE =** parameter in the last line of the below code block.

```
.....
;

<<Placeholder1: Paste the PUBCLUSTER Macro Definition Here >>

DATA&EM_EXPORT_TRAIN;
       SET&EM_IMPORT_DATA;


<<Placeholder 2: Paste the Score Code from the Cluster Node >>


%PROBCLUSTER(CLUSVads,PARRAY_D,PARRAY_ED,SIZE=3);

RUN;
```

C. CUSTOMER OPINION SURVEY (XYZ COMPANY)

How important are the following issues to you in choosing a supplier for hydraulic, pneumatic and related products?	Not at all important								Extremely important
1. The reliability of the supplier (reliab)	1	2	3	4	5	6	7	8	9
2. The timeliness of the deliveries by the supplier (time)	1	2	3	4	5	6	7	8	9
3. The availability of a large breadth of products to choose from (av_br)	1	2	3	4	5	6	7	8	9
4. The availability of well documented technical specification (av_spec)	1	2	3	4	5	6	7	8	9
5. The price of products (price)	1	2	3	4	5	6	7	8	9
6. The credit policy of the supplier (credit)	1	2	3	4	5	6	7	8	9
7. The availability of electronic payment/debit option (av_pay)	1	2	3	4	5	6	7	8	9
8. The return policy of the supplier (return)	1	2	3	4	5	6	7	8	9
9. The warranty coverage provided by the supplier (warranty)	1	2	3	4	5	6	7	8	9
10 The ability to talk directly to a salesperson about your needs (talk_dir)	1	2	3	4	5	6	7	8	9

REFERENCES

Blattberg, R., Kim, B. & Neslin, S. (2008). *Database Marketing: Analyzing and Managing Customers*. New York: Springer Science

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

Budayan, C., Dikmen, I. & Birgonul, T. (2008). Comparing the performance of traditional cluster analysis, self organizing maps and fuzzy C – means method for strategic grouping. *Expert Systems with Applications*, 36, pp 11772 – 11781

Chuang, K., Chiu, M., Lin, C., & Chen, J. (1999). Model free functional MRI analysis using Kohonen clustering neural network and Fuzzy C Means. *IEEE Transactions on Medical Imaging*, 18(12), pp 1117 – 1128

Clarke, A. (2009). Bridging industrial segmentation theory and practice. *Journal of Business-to-Business Marketing*, 16(4), pp 343 – 373

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

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

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.

Ozer, M. (2001). User segmentation of online music services using fuzzy clustering. *Omega Int J of Management Science*, 29, pp 193 – 206

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

ACKNOWLEDGEMENTS

Our thanks to Dr. Don Wedding of SAS® for pointing out the distance array available in the k-means score node for developing our macro.

TRADEMARKS

CONTACT INFORMATION

Dipanjan Kumar Dey, Email: dipanjan.dey@gmail.com

Satish Garla, Email: satish.garla@okstate.edu

Goutam Chakraborty, Email: goutam.chakraborty@okstate.edu

Brief Bios:

Dipanjan Kumar Dey is a visiting research scholar to Oklahoma State University, Spears School of Business from IBS, Hyderabad, India. He has professional experience as a sales manager. He has interests in business applications of Software packages. His research interests are in the field of market segmentation techniques especially application of predictive modeling in market segmentation using SAS®.

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 Business Consultant. He is SAS® Certified Advanced Programmer for SAS 9® and Certified Predictive Modeler Using SAS Enterprise Miner 6®.

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®.