Paper 3150-2015

Statistical Evaluation of the Doughnut Clustering Method for Product Affinity Segmentation

Juan Ma[1], Darius Baer[2], Goutam Chakraborty[1]

[1]Oklahoma State University, Stillwater, OK [2]SAS Institute Inc., Cary, NC

ABSTRACT

Product affinity segmentation is a powerful technique for marketers and sales professionals to gain a good understanding of customers' needs, preferences, and purchase behavior. Performing product affinity segmentation is quite challenging in practice because product level data usually have high skewness, high kurtosis, and large percentage of zero values. The Doughnut clustering method has been shown to be effective using real data, and was presented at SAS Global Forum 2013 (Baer & Chakraborty, 2013). However, the Doughnut clustering method is not a panacea for addressing the product affinity segmentation problem. There is a clear need for a comprehensive evaluation of this method in order to be able to develop generic guidelines for practitioners on when to apply the method. In this paper, we meet the need by evaluating the Doughnut clustering method on simulated data with different levels of skewness, kurtosis, and percentage of zero values. We developed a five-step approach based on Fleishman's power method to generate synthetic data with prescribed parameters. Subsequently, we designed and conducted a set of experiments to apply the Doughnut clustering method as well as the traditional K-means method as benchmark on the simulated data. We draw conclusions on the performance of the Doughnut clustering method by comparing the clustering validity metric "the ratio of between-cluster variance to within-cluster variance" as well as the relative proportion of cluster sizes against those of K-means. In certain data situations, the Doughnut clustering method is shown to produce an acceptable clustering solution when other approaches fail.

INTRODUCTION

It is critical for marketers and sales professionals to gain good understanding of customers' needs, preferences, and purchase behavior so that customers receive the most relevant marketing message and promotion and companies obtain increased revenue and/or profit. To serve this goal, one can generally conduct *market basket analysis* or *product affinity segmentation* (Collica, 2011). A typical example of market basket analysis is to analyze customer transactions at point-of-sale (POS) to discover association rules, i.e., which products go together in a customer's shopping cart. With the association rules, companies can arrange their items on the supermarket shelves so that products purchased together will be seen together by customers. On the other hand, product affinity segmentation is to cluster customers into similar groups that share common purchase patterns. With the product affinity segments, marketers and sales professionals can provide customized offers and promotions to members in each segment. Both market basket analysis and product affinity segmentation can be employed to identify cross-selling and up-selling opportunities and to promote appealing packages of products and services.

Between the market basket analysis and product affinity segmentation, the latter "provides the most powerful and profitable information for marketing campaigns and communication with customers (Baer & Chakraborty, 2013)", which is our focus in this paper. Although product affinity is conceptually attractive and can effectively aid in campaign and promotion success, it is quite difficult to perform in practice. The challenges lie in two aspects. i) Transactional data used for product affinity segmentation whether measured in quantities of products or revenue usually have very high skewness and kurtosis. This is because it is often the case that a few customers buy a lot and the majority buys very little. ii) Many transactional variables (product quantities) have a significant percentage of zero values due to the fact that moderate-sized and/or large-sized companies generally have a variety of products and customers usually only buy a selected number of what's appealing to them. Due to these difficulties of transactional data, readily available clustering algorithms often result in disproportionately sized segments such as a single overly dominating segment along with many tiny segments. In this paper, we extend the research

work in (Baer, 2012; Baer & Chakraborty, 2013) by systematically evaluating the performance of Doughnut clustering method on simulated data.

WHY ARE WE EVALUATING THE DOUGHNUT CLUSTERING METHOD?

The Doughnut clustering method, first presented in (Baer, 2012), was designed for product affinity segmentation on product level data with high skewness, high kurtosis, and large percentage of zero values. In such a case, using standard K-means clustering method, one often ends up with a number of clusters in the center of the dimensional space because there is a higher density of observations toward the center. The core idea behind the Doughnut clustering method is to enforce a single cluster at the center of the dimensional space and all other clusters to be appropriately positioned in other areas. This is to make sure that each cluster has a sufficient proportion of the overall population and the differences among the clusters are significant and meaningful. Although the Doughnut clustering method has been used in a wide variety of industries, and an application example was demonstrated in (Baer & Chakraborty, 2013) from soup to nuts, little has been done to evaluate the Doughnut clustering method from an academic point of view. There is a clear need for such evaluation for two reasons. First, although the Doughnut clustering method has been shown to yield satisfactory clustering results in some applications, the effectiveness of this method alone is unclear because there are some other ad-hoc techniques (i.e., Softmax transformation) applied at the same time to aid in clustering based on problem specific characteristics such as high skewness of the data. Second, we know the Doughnut clustering method is not a panacea for product affinity segmentation by itself. And, quantitative guidelines on when to use the Doughnut clustering method are not known. As an initial step in evaluating the Doughnut clustering method in academic settings, we hope to shed lights on its effectiveness and develop clear guidelines to marketers and other analytics professionals.

SIMULATED DATA GENERATION METHODOLOGY

In this section, we describe a five-step approach to generate simulated data with given kurtosis, skewness, and percentage of zero values. The approach is partially based on Fleishman's cubic transformation method (Fleishman, 1978) to generate non-normal distribution with given skewness and kurtosis. To make this paper self-contained, we briefly recap Fleishman's method first.

In the literature, a variety of methods has been developed to generate non-normally distributed variables with prescribed degrees of skewness and kurtosis. The interested reader is referred to (Reinartz, Echambadi, & Chin, 2002) for detailed information. All the methods follow some general criteria: a) the input has parameters characterizing the distribution; b) one can change the distribution by changing the input parameters with minimum difficulty; c) the input should be able to characterize different distribution with various deviations from normality; d) the data generation process can be operated efficiently. Based on these criteria, Fleishman's method uses a polynomial transformation (thus also noted as power method) of the following form:

$$\mathrm{Y} = \mathrm{a} + \mathrm{bX} + \mathrm{cX}^2 + \mathrm{dX}^3 \quad (1)$$

where X is normally distributed with zero mean and unit variance, i.e. $N(0,1)$. The constants $a, b, c,$ and d may be chosen such that Y has a distribution with specified moments of the first four moments, i.e., mean ($E[Y]$), variance ($E[Y^2]$), skewness ($E[Y^3]$), and kurtosis ($E[Y^4]$). For the sake of convenience, let $E[Y] = 0$ and $E[Y^2] = 1$. Since $E[Y] = a + c$, it follows $a = -c$. We further have the following Fleishman's equations.

$$E[Y^2] = b^2 + 6bd + 2c^2 + 15d^2 \quad (2)$$

$$E[Y^3] = 2c(b^2 + 24bd + 105d^2 + 2) \quad (3)$$

$$E[Y^4] = 24(bd + c^2[1 + b^2 + 28bd] + d^2[12 + 48bd + 141c^2 + 225d^2]) \quad (4)$$

Therefore, in order to generate non-normally distributed variable Y with given skewness ($E[Y^3]$) and kurtosis ($E[Y^4]$), one just needs to solve the Fleishman's equation system of Eq. (2)-(4). It is worth noting that the *Fleishman method will not work for all combinations of skewness and kurtosis*; the equation system may not have a real number solution. The function that implements Fleishman's method is now available in SAS/IML. The interested reader is referred to (Wicklin, 2013) for implementation details.

Since we also want to control the percentage of zero values in addition to skewness and kurtosis of the simulated data, a five-step approach based on Fleishman's method is developed as shown in Table 1.

Input: sample size $\boldsymbol{N}$, skewness $\boldsymbol{E[Y^3]}$, kurtosis $\boldsymbol{E[Y^4]}$, percentage of zero values $\boldsymbol{\alpha}$, mean (optional), standard deviation (optional)

1. Use the Fleishman equations to get the values of $\boldsymbol{a, b, c}$ and $\boldsymbol{d}$ for a desired feasible combination of skewness and kurtosis.
2. Generate $\boldsymbol{N}$ data points from normally distributed variables $\boldsymbol{X}$ with mean $\boldsymbol{E[X] = 0}$ and standard deviation $\boldsymbol{\sigma = 1}$.
3. Transform the data in step 2 with the Fleishman coefficients obtained in step 1 to those with given skewness and kurtosis following Eq. (1).
4. (*Optional*) Transform the data from step 3 to the desired mean and standard deviation, if any. New data = desired mean + (data from step 3)* desired standard deviation.
5. This step assigns a user-specified percentage, denoted by $\boldsymbol{\alpha} \in (\boldsymbol{0, 1})$, of randomly selected data points to be zero. Give the data from step 4,

   For data point $\boldsymbol{1, \ldots, N}$, repeat

   - Sample a data point from a variable uniformly distributed variable between 0 and 1.
   - If the data value is less than the user-specified percentage $\boldsymbol{\alpha}$, let the data point from step 4 to be zero; otherwise keep the data point from step 4 unchanged.

   End-repeat

Table 1: The Five-Step Data Generation Approach

The generated data from Step 3 will have the desired skewness and kurtosis. Moreover, output from step 4 will have the desired mean, standard deviation, skewness, and kurtosis if the optional step 4 is executed. Additionally, mean, standard deviation, skewness, and kurtosis will change after step 5. The magnitude of the changes depends on the percentages prescribed by the specific user, i.e. input parameter α. Larger α tends to result in bigger jumps in mean, standard deviation, skewness and kurtosis.

Note that Fleishman's method assumes non-correlated variables. A more recent study by Vale and Maurelli (Vale & Maurelli, 1983) extended Fleishman's method to generate multivariate data with a prescribed correlation. The implemented function of this Vale-Maurelli method is also available in SAS/IML (Wicklin, 2013). As our five-step approach is based on Fleishman's method, we are implicitly assuming the variables are non-correlated. However, if we replace step 1 with the Vale-Maurelli method, we will be able to generate correlated multivariate non-normal data. In that case, the degree of correlation will also be affected by step 5 in a way similar to skewness and kurtosis. As an initial study focused the Doughnut clustering method for product affinity segmentation on simulated data, we do not consider correlations between transaction variables in this paper. However, the unique interaction between percentages of zero values and skewness/kurtosis/correlation can be an interesting direction for future research.

IMPLEMENTATION OF THE DATA GENERATION METHODOLOGY

We introduce the key points of our implementation of the five-step approach in this subsection. For step 1, we employed the "*Solve"* function in Mathematica® to solve the Fleishman's equations for a given feasible combination of skewness and kurtosis. For instance, given skewness = 1.5 and kurtosis = 25, by solving the corresponding Fleishman's nonlinear equation system we have a = -0.201827, b = 1.50632, c = 0.201827, d = -0.328259.

In accordance to the real data employed in (Baer & Chakraborty, 2013), we create simulated transaction data with eight product classes and 30,000 entries. For each product class (i.e. A, B, C, D, E, F, G, and

H), the number of units purchased is recorded. Under this setting, we implemented step 2 using the function "*rand*" as follows.

```
Data work.SimulatedData;
    do ID=1 to &numRecords;
           *generate random product quantity;
           Prod_Quant_A=rand("normal", 0, 1);
           Prod_Quant_B=rand("normal", 0, 1);
           Prod_Quant_C=rand("normal", 0, 1);
           Prod_Quant_D=rand("normal", 0, 1);
           Prod_Quant_E=rand("normal", 0, 1);
           Prod_Quant_F=rand("normal", 0, 1);
           Prod_Quant_G=rand("normal", 0, 1);
           Prod_Quant_H=rand("normal", 0, 1);
           output;
    end;
run;
```

Step 3 is to convert the generated the data from step 2 into non-normal data with the coefficients obtained in step1. We show the implementation below.

```
Data work.SimulatedData;
set work.SimulatedData;
    Prod_Quant_A_=&a + &b*Prod_Quant_A + &c*Prod_Quant_A**2 + &d*Prod_Quant_A**3;
    Prod_Quant_B_=&a + &b*Prod_Quant_B + &c*Prod_Quant_B**2 + &d*Prod_Quant_B**3;
    Prod_Quant_C_=&a + &b*Prod_Quant_C + &c*Prod_Quant_C**2 + &d*Prod_Quant_C**3;
    Prod_Quant_D_=&a + &b*Prod_Quant_D + &c*Prod_Quant_D**2 + &d*Prod_Quant_D**3;
    Prod_Quant_E_=&a + &b*Prod_Quant_E + &c*Prod_Quant_E**2 + &d*Prod_Quant_E**3;
    Prod_Quant_F_=&a + &b*Prod_Quant_F + &c*Prod_Quant_F**2 + &d*Prod_Quant_F**3;
    Prod_Quant_G_=&a + &b*Prod_Quant_G + &c*Prod_Quant_G**2 + &d*Prod_Quant_G**3;
    Prod_Quant_H_=&a + &b*Prod_Quant_H + &c*Prod_Quant_H**2 + &d*Prod_Quant_H**3;
run;
```

The last step is to assign a prescribed percentage of zero values (i.e. input parameter α) as follows. The value of α is stored in the MACRO variable *"pct"*. We assume each product class variable has the same percentage of zero values.

```
Data work.SimulatedData;
set work.SimulatedData;
    i1=rand("UNIFORM"); if i1<=&pct then Prod_Quant_A=0;
    i2=rand("UNIFORM"); if i2<=&pct then Prod_Quant_B=0;
    i3=rand("UNIFORM"); if i3<=&pct then Prod_Quant_C=0;
    i4=rand("UNIFORM"); if i4<=&pct then Prod_Quant_D=0;
    i5=rand("UNIFORM"); if i5<=&pct then Prod_Quant_E=0;
    i6=rand("UNIFORM"); if i6<=&pct then Prod_Quant_F=0;
    i7=rand("UNIFORM"); if i7<=&pct then Prod_Quant_G=0;
    i8=rand("UNIFORM"); if i8<=&pct then Prod_Quant_H=0;
run;
```

Due to length limitation, we omit the description of the Doughnut clustering method implementation in this paper. The interested is strongly encouraged to read (Baer & Chakraborty, 2013) for a complete description.

EXPERIMENT DESIGN

In order to generate a meaningful test bed for the Doughnut clustering method, we arrange the input for the five-step data generation approach as follows. We consider two scales of skewness and kurtosis respectively; namely skewness=0 or 5.75 and kurtosis =0 or 95.75. As our initial exploration results show that the final clustering results are more sensitive to the percentage of zero values, we consider 7 scales of percentage values, i.e., $\alpha = 0.0, 0.3, 0.4, 0.5, 0.6, 0.7,$ and 0.8. In addition, because skewness= 5.75 and kurtosis=0 is not a feasible combination for Fleishman's equation system, we ended up with 21 combinations of feasible input parameters for the five-step approach as shown in Table 2. Each ID

"XX.0X" in the table should be interpreted as follows. The first letter "X" stands for skewness level, i.e., either Low (0) or High (5.75); the second letter "X" stands for kurtosis level, i.e., either Low (0) or High (95.75); the last letter "X" stands for the value of parameter α.

ID	Skewness	Kurtosis	Pcnt of 0s	ID	Skewness	Kurtosis	Pcnt of 0s	ID	Skewness	Kurtosis	Pcnt of 0s
LL0.0	0	0	0	LH0.0	0	95.75	0	HH0.0	5.75	95.75	0
LL0.3	0	0	0.3	LH0.3	0	95.75	0.3	HH0.3	5.75	95.75	0.3
LL0.4	0	0	0.4	LH0.4	0	95.75	0.4	HH0.4	5.75	95.75	0.4
LL0.5	0	0	0.5	LH0.5	0	95.75	0.5	HH0.5	5.75	95.75	0.5
LL0.6	0	0	0.6	LH0.6	0	95.75	0.6	HH0.6	5.75	95.75	0.6
LL0.7	0	0	0.7	LH0.7	0	95.75	0.7	HH0.7	5.75	95.75	0.7
LL0.8	0	0	0.8	LH0.8	0	95.75	0.8	HH0.8	5.75	95.75	0.8

Table 2: Input Parameters for the Five-Step Data Generation Approach

We note that the corresponding skewness and kurtosis of the final simulated data is not the same as input skewness and kurtosis values except for the case $\alpha = 0$. This is due to step 5 in the five-step approach as explained earlier. For instance, the final degrees of skewness and kurtosis of simulated data IDs HH0.3 and HH0.8 are as shown in Table 3 and Table 4.

Variable	Skewness	Kurtosis	Pcnt of 0s
Prod_Quant_A	7.16	135.33	30.23%
Prod_Quant_B	7.48	143.89	30.51%
Prod_Quant_C	7.05	132.86	30.89%
Prod_Quant_D	7.99	156.26	30.84%
Prod_Quant_E	7.09	134.34	30.54%
Prod_Quant_F	7.12	113.95	30.88%
Prod_Quant_G	5.76	116.32	31.01%
Prod_Quant_H	8.18	141.42	30.71%

Table 3: The Final Skewness and Kurtosis Values for Experiment ID HH0.3

Variable	Skewness	Kurtosis	Pcnt of 0s
Prod_Quant_A	10.87	284.29	80.78%
Prod_Quant_B	12.26	304.16	80.24%
Prod_Quant_C	14.93	437.09	80.32%
Prod_Quant_D	15.51	525.22	80.36%
Prod_Quant_E	13.77	396.59	80.02%
Prod_Quant_F	8.65	258.14	80.23%
Prod_Quant_G	12.97	475.97	79.77%
Prod_Quant_H	10.75	266.24	80.22%

Table 4: The Final Skewness and Kurtosis Values for Experiment ID HH0.8

RESULTS

With experiments set up as described in the previous section, we run both the Doughnut clustering method and the traditional K-means method. Because standardization and a cap of 3 standard deviations are commonly used in practice for product affinity segmentation, we apply these two data manipulation techniques on the simulated data before feeding them to both K-means and the Doughnut clustering method.

When it comes to evaluating clustering results, one generally has to employ appropriate criteria and techniques. Such evaluation process is called *cluster validity* and the criteria or techniques are named *clustering validity metrics.* Moreover, clustering validity is formally defined as "procedures that evaluate the results of clustering analysis in a quantitative and objective fashion" (Jain & Dubes, 1988). Meanwhile, clustering validity metrics measure the merit of clustering results in a quantitative manner. Commonly used clustering validity metrics generally fall under three categories (Halkidi, Batistakis, & Vazirgiannis, 2001): i) external criteria where we evaluate the results based on pre-specified structure imposed on a data set; ii) internal criteria where we evaluate the results in terms of quantities that invoke the vectors of

the data themselves, e.g., proximity matrix; and iii) relative criteria where the evaluation is conducted by comparing the structure against other clustering schemes. Some of the commonly used metrics include homogeneity and separation, Silhouette Coefficient (Rousseeuw, 1987), cluster sizes and consistency, and many others.

In this paper we employ *cluster size* and *the ratio of between-cluster variance to within-cluster variance* as validity metrics. Cluster size, classified as external criteria, is used to check if the size distribution across segments is reasonable for downstream marketing strategy application from a managerial level. An overly dominating cluster with many tiny clusters is clearly not preferable. The internal criteria, ratio of between-cluster variance to within-cluster variance, provided as part of the statistics output of PROC FASTLCUS denoted by RSQ_ratios, measures the ratio of difference between groups divided by homogeneity within groups. Therefore, the larger the RSQ_ratio is, the better the clustering results are.

We report the cluster sizes in terms of percentages (cluster frequency divided by overall frequency multiplied by 100) for generated data in Table 5, Table 6, and Table 7. Correspondingly, the RSQ_ratio values are reported in Figure 1, Figure 2, and Figure 3.

α	0		0.3		0.4		0.5		0.6		0.7		0.8	
Cluster	**K**	**D**	**K**	**D**	**K**	**D**	**K**	**D**	**K**	**D**	**K**	**D**	**K**	**D**
1	11.2	1.6	10.0	4.2	10.1	5.5	9.7	8.7	7.2	11.1	7.0	17.9	**4.6**	**29.2**
2	10.9	12.2	9.9	11.2	9.0	11.3	8.4	9.3	8.4	9.5	5.9	10.1	**5.3**	**7.9**
3	11.0	12.4	10.5	12.6	22.9	11.2	8.9	10.1	37.8	10.6	6.3	7.8	**5.4**	**9.3**
4	11.1	12.3	10.4	12.8	9.9	11.8	9.2	10.9	8.3	9.9	9.3	12.1	**7.4**	**8.4**
5	11.5	12.6	10.2	12.0	9.9	11.5	8.4	12.5	8.3	11.9	7.0	8.4	**5.2**	**16.4**
6	11.1	12.1	10.7	12.8	9.3	12.4	29.7	10.9	8.2	12.6	6.2	10.4	**5.6**	**7.2**
7	11.1	12.3	10.2	12.2	9.9	13.3	9.3	11.7	7.3	12.3	41.6	10.7	**49.8**	**6.5**
8	11.4	12.4	10.5	11.8	9.9	11.9	8.0	13.4	7.4	11.5	9.9	8.5	**7.9**	**9.0**
9	10.8	12.2	17.7	10.5	8.9	11.1	8.3	12.6	7.2	10.7	6.8	14.0	**8.9**	**6.1**

K: K-means D: Doughnut cluster method

Table 5: Resulted Cluster Sizes in Terms of Percentages from K-means and Doughnut Clustering Method on Simulated Data with Skewness =0 and Kurtosis = 0

α	0		0.3		0.4		0.5		0.6		0.7		0.8	
Cluster	**K**	**D**	**K**	**D**	**K**	**D**	**K**	**D**	**K**	**D**	**K**	**D**	**K**	**D**
1	1.6	1.3	1.3	4.6	1.2	5.9	10.1	7.1	0.8	11.1	6.2	17.3	**4.8**	**26.8**
2	11.4	12.3	14.2	11.4	1.3	10.5	11.6	12.0	10.0	11.7	8.4	11.2	**5.9**	**7.2**
3	13.2	12.5	17.8	11.1	15.7	12.3	11.9	9.5	8.1	11.6	7.9	10.9	**6.4**	**3.4**
4	15.5	12.6	14.6	10.8	1.2	12.0	10.3	12.8	27.8	13.0	10.3	10.0	**5.9**	**22.4**
5	14.4	12.6	15.9	12.3	23.0	12.3	11.9	12.2	13.0	8.8	8.0	9.2	**45.4**	**8.0**
6	1.8	12.5	16.3	13.4	16.2	11.1	11.3	13.4	9.0	10.0	7.3	10.4	**10.0**	**11.0**
7	13.1	12.1	16.8	11.8	14.5	12.7	1.1	9.8	8.8	12.1	36.3	9.6	**8.1**	**7.5**
8	14.2	11.9	1.5	12.4	15.3	11.0	9.5	10.8	12.7	11.4	6.9	11.6	**5.9**	**7.8**
9	14.8	12.3	1.5	12.2	11.5	12.3	22.4	12.3	9.7	10.2	8.7	9.6	**7.5**	**5.9**

K: K-means D: Doughnut cluster method

Table 6 Resulted Cluster Sizes in Terms of Percentages from K-means and Doughnut Clustering Method on Simulated Data with Skewness =0 and Kurtosis =95.75

α	0		0.3		0.4		0.5		0.6		0.7		0.8	
Cluster	**K**	**D**	**K**	**D**	**K**	**D**	**K**	**D**	**K**	**D**	**K**	**D**	**K**	**D**
1	29.0	6.4	2.7	7.5	2.5	10.8	2.3	11.1	1.8	15.3	6.7	22.2	**4.8**	**26.5**
2	3.5	8.7	37.0	11.6	2.6	13.6	72.0	12.8	1.9	8.5	6.8	8.7	**1.0**	**6.5**
3	3.4	12.5	2.9	12.0	2.5	9.8	2.2	13.2	2.0	11.5	6.6	8.3	**61.7**	**6.7**
4	3.6	10.8	38.9	11.7	40.0	11.8	2.3	7.3	10.8	10.8	1.4	9.7	**4.9**	**9.9**
5	24.0	13.0	7.5	11.9	38.0	12.5	12.1	10.1	68.8	8.5	5.9	8.2	**4.9**	**24.4**
6	3.3	13.0	2.8	11.5	2.4	13.2	2.4	13.6	9.1	9.2	6.7	10.2	**8.5**	**6.9**
7	19.0	10.7	2.8	10.7	2.6	13.2	2.2	11.3	1.8	14.1	6.3	15.3	**4.9**	**7.0**
8	3.2	13.0	2.8	11.3	6.9	15.1	2.2	9.2	1.8	11.6	53.0	10.0	**4.4**	**5.7**
9	11.0	12.0	2.7	11.8	2.5	0.0	2.3	11.6	2.0	10.6	6.6	7.4	**5.0**	**6.3**

K: K-means D: Doughnut cluster method

Table 7: Resulted Cluster Sizes in Terms of Percentages from K-means and Doughnut Clustering Method on Simulated with Skewness =5.75 and Kurtosis = 95.75

HOW EVENLY DISTRIBUTED ARE CLUSTER SIZES?

As shown in Table 5, Table 6, and Table 7, the Doughnut clustering method brings more evenly distributed segments compared to the traditional K-means method when simulated data have larger skewness, kurtosis, and percentage of zero values. For example, as shown in Table 7, when the percentage of zero values $\alpha = 0.8$, we can see that, while K-means results in one cluster with 61.7% of the total data points and eight tiny segments, Doughnut method yields two moderate size clusters (26.5% and 24.4% respectively) and seven relatively smaller segments.

Since it is critical in business practice to have evenly distributed cluster sizes, a metric to quantify such a quality becomes desirable. In this paper we use the p-value for Chi-square test to achieve this goal. We assume an evenly distributed 9-cluster solution attains a size distribution of 12%, 11%, 11%, 11%, 11%, 11%, 11%, 11%, and 11%. For each of the 21 simulated data, we next compare the size distribution of results from both the Doughnut clustering method and K-means against this "standard even distribution" using Chi-square test. Thus, the lower the p-value for Chi-square test is, the less likely the corresponding solution is similarly (evenly) distributed. Specifically, a perfect agreement would result in a p-value of 1. We report p-values on all 21 simulated data in Table 8 below.

Input Skewness=0, Kurtosis=0							
Pnct of zero values	**0**	**0.3**	0.4	0.5	0.6	0.7	0.8
K-Means	1	0.9831	0.7446	0.2061	0.0114	0.0014	<.0001
Doughnut	0.3745	0.8251	0.9438	0.9968	0.9999	0.9441	0.1212
Input Skewness=0, Kurtosis=95.75							
Pnct of zero values	0	0.3	0.4	0.5	0.6	0.7	0.8
K-Means	0.0284	0.0004	0.0001	0.1322	0.0188	0.0167	0.0001
Doughnut	0.317	0.8741	0.9626	0.981	0.9997	0.9943	0.0245
Input Skewness=5.75, Kurtosis=95.75							
Pnct of zero values	0	0.3	0.4	0.5	0.6	0.7	0.8
K-Means	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Doughnut	0.9549	0.9965	0.1266	0.9914	0.9908	0.6866	0.0244

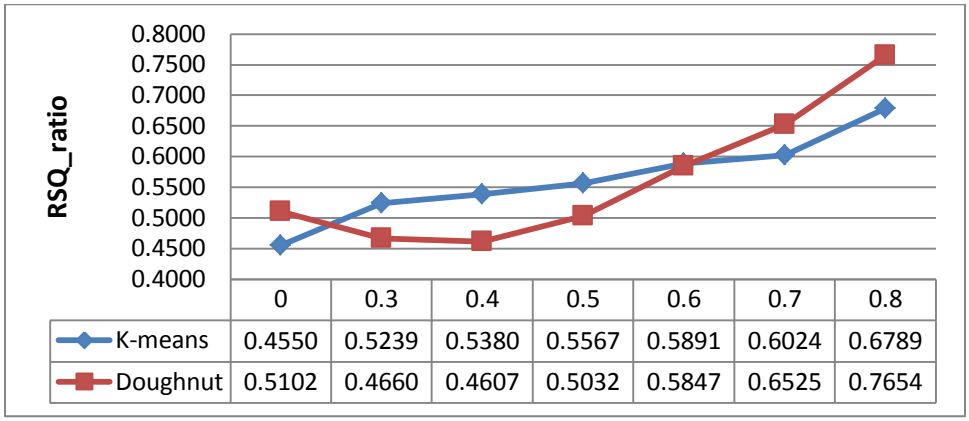
Table 8: The p-Values of Chi-square Tests for Cluster Size Distributions on Different Simulated Data

It is clear from Table 8 that the p-values of results from Doughnut clustering method is larger in almost all the cases except for two special cases where Skewness=0, Kurtosis=0, and percentage of zero values =0 or 0.3. In other words, results from the Doughnut clustering method is more evenly distributed than those from K-means method on most of our tested data. This exciting observation signifies the benefits of the Doughnut clustering method in business applications.

WHAT IS RATIO OF BETWEEN-CLUSTER VARIANCE TO WITHIN-CLUSTER VARIANCE?

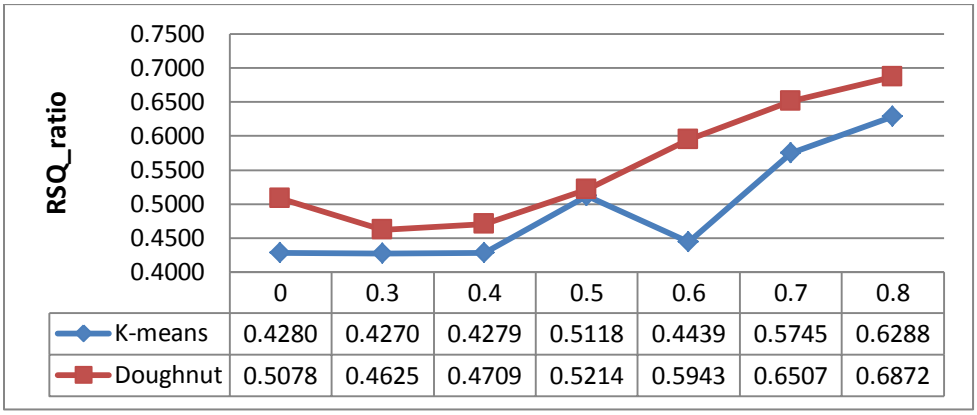
By comparing RSQ_ratio values in Figure 1, Figure 2, and Figure 3, we can see that the Doughnut clustering method is obviously performing better than K-means when data come with small skewness and

high kurtosis regardless of the values of percentages of zeros in the data. On the other hand, when simulated data has skewness of 5.75 and kurtosis of 95.75, the Doughnut clustering method is only better when the percentage of zeros is high, i.e. $\alpha = 0.8$. In other tested cases, the advantages of Doughnut clustering method over the traditional K-means method are not pronounced.



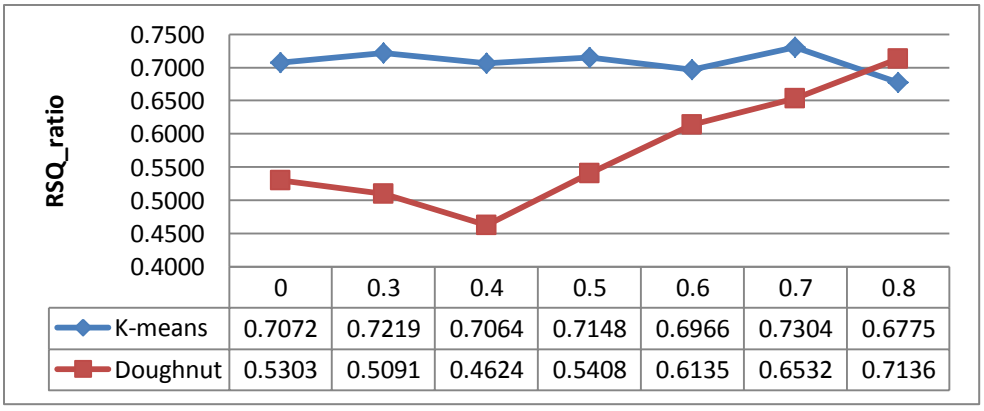
	0	0.3	0.4	0.5	0.6	0.7	0.8
K-means	0.4550	0.5239	0.5380	0.5567	0.5891	0.6024	0.6789
Doughnut	0.5102	0.4660	0.4607	0.5032	0.5847	0.6525	0.7654

Figure 1: Comparison of RSQ_ratio Values of K-means and Doughnut for Different Level of Percentages of Zero Values While Fixing Skewness =0 and Kurtosis =0



	0	0.3	0.4	0.5	0.6	0.7	0.8
K-means	0.4280	0.4270	0.4279	0.5118	0.4439	0.5745	0.6288
Doughnut	0.5078	0.4625	0.4709	0.5214	0.5943	0.6507	0.6872

Figure 2: Comparison of RSQ_ratio Values of K-means and Doughnut for Different Level of Percentages of Zero Values While Fixing Skewness =0 and Kurtosis =95.75



	0	0.3	0.4	0.5	0.6	0.7	0.8
K-means	0.7072	0.7219	0.7064	0.7148	0.6966	0.7304	0.6775
Doughnut	0.5303	0.5091	0.4624	0.5408	0.6135	0.6532	0.7136

Figure 3: Comparison of RSQ_ratio Values of K-means and Doughnut for Different Level of Percentages of Zero Values While Fixing Skewness =5.75 and Kurtosis =95.75

CLUSTER DIFFERENTIATION BASED ON DOUGHNUT CLUSTERING METHOD

In order to examine cluster differentiation to better understand customer purchase behavior in different groups, we calculate the average number of each product purchased within each cluster. Experiments were conducted separately on clusters generated by both K-means and the Doughnut clustering method for all the 21 simulated data. As an example, we present the results for the simulated data with input skewness= 5.75, kurtosis= 95.75, and percent of zero values= 0.8. The means of different variables within clusters generated by K-means are reported in Table 9 and Figure 4. Table 10 and Figure 5 present the means of different variables within clusters generated by the Doughnut clustering method. Since the clustering was done to keep customers with similar purchase patterns together, distinct patterns can be observed from the means presented.

As shown in Table 9, Table 10, Figure 4, and Figure 5, both the K-means and the Doughnut clustering methods result in clusters that have high means for only one product. This is to be expected and the results can be satisfactorily used for product marketing. As a point of clarification, the reader will note that many of the mean values are less than zero. This seems strange from a customer purchase perspective as no customer would purchase less than zero. However, as an academic exercise in applying the Fleishman's method, we obtained many values that were less than zero. Had we transformed all the negative values to zero, we would have changed the desired Skewness, Kurtosis, and percent of zero values. Thus, in this paper we keep the negative data points in order to gain more accurate control over the three parameters.

While the product affinity is important and having high means for only one product is good, the size of the cluster cannot be overlooked. Therefore, it is important that the clusters be as evenly sized as possible. The size of the cluster is the clear differentiator between the K-means and the Doughnut methods.

Variable Mean Within Each Cluster											
Cluster	**Freq**	**Pcnt**	**All**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
All	30000	100.0%	5.9	0.8	0.7	0.8	0.9	0.8	0.7	0.6	0.8
1	1431	4.8%	16.0	-0.9	-0.9	-1.4	-1.0	-1.1	**24.8**	-1.6	-1.8
2	304	1.0%	137.5	-0.5	0.4	2.5	0.7	-0.5	**134.6**	0.7	-0.2
3	18522	61.7%	-17.0	-2.1	-2.1	-2.0	-2.0	-2.1	-2.3	-2.2	-2.1
4	1465	4.9%	45.3	-0.6	-1.2	-1.6	-0.7	**51.6**	-1.0	-0.7	-0.6
5	1461	4.9%	45.6	-0.1	**50.8**	-1.1	-0.6	-1.0	-1.1	-0.7	-0.8
6	2552	8.5%	42.7	**26.4**	-1.8	-1.5	-1.3	-1.7	-1.6	**25.7**	-1.4
7	1456	4.9%	42.2	-1.1	-1.2	**50.6**	-1.2	-1.1	-1.6	-0.8	-1.3
8	1310	4.4%	47.2	-1.0	-1.7	-1.0	**56.0**	-1.4	-1.2	-1.1	-1.4
9	1499	5.0%	41.7	-0.4	-1.6	-1.1	-1.1	-1.1	-1.7	-0.8	**49.5**

Table 9: Results from K-means Method for Simulated Data with Skewness =5.75, Kurtosis =95.75, and Percentage of zero values = 0.8

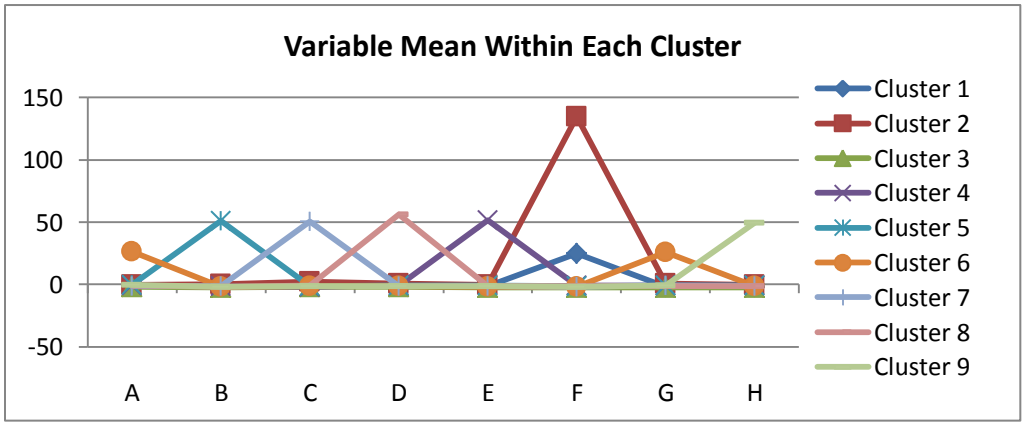


Figure 4: Results from K-means Method for Simulated Data with Skewness =5.75, Kurtosis =95.75 and Percentage of zero values = 0.8

Variable Mean Within Each Cluster											
Cluster	**Freq**	**Pcnt**	**All**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
All	30000	100.0%	5.9	0.8	0.7	0.8	0.9	0.8	0.7	0.6	0.8
1	7959	26.5%	-3.0	-0.1	0.0	-0.1	-0.2	-1.9	-0.2	-0.1	-0.3
2	1961	6.5%	35.5	**38.7**	1.0	-0.1	-1.6	-0.4	-0.6	1.0	-2.6
3	2019	6.7%	35.6	0.4	0.7	0.5	-2.2	-2.1	-0.8	1.1	**38.1**
4	2972	9.9%	-6.8	-2.7	-12.6	**19.4**	-2.1	-2.3	-2.3	-2.4	-1.9
5	7331	24.4%	-24.6	-4.5	**6.7**	-4.6	-4.8	-2.9	-4.9	-5.0	-4.6
6	2058	6.9%	35.4	-1.6	0.0	-0.6	-2.0	**39.0**	0.2	-0.1	0.5
7	2102	7.0%	42.9	-0.1	0.9	0.1	**41.9**	-0.2	0.1	0.0	0.2
8	1723	5.7%	24.9	-2.1	0.8	0.1	-2.4	-2.3	-1.8	**35.0**	-2.5
9	1875	6.3%	29.6	-2.1	1.1	0.7	-2.0	-2.0	**37.1**	-0.7	-2.6

Table 10: Results from the Doughnut Clustering Method for Simulated Data with Skewness =5.75, Kurtosis =95.75, and Percentage of zero values = 0.8

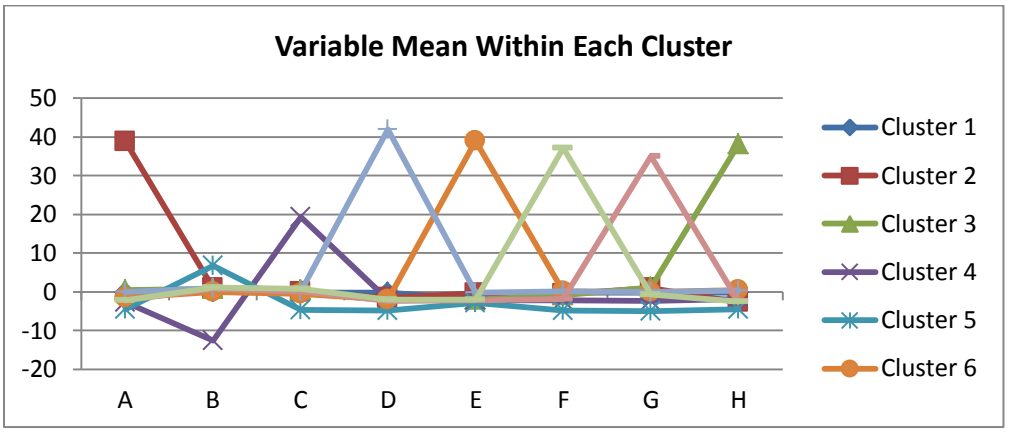


Figure 5: Results from the Doughnut Clustering Method for Simulated Data with Skewness =5.75, Kurtosis =95.75, and Percentage of zero values = 0.8

CONCLUSION

In this paper, we evaluated the performance of the Doughnut clustering method for product affinity problem by conducting a comprehensive numerical study on simulated data with different levels of skewness, kurtosis, and percentages of zero values. Our results show that the Doughnut method works better than the traditional K-means method on simulated data with large kurtosis and large percentage of zero values. This observation emphasizes the advantages of the Doughnut clustering method over tradition K-means method because reducing skewness by transformation methods (e.g. Softmax transformation) is practically more achievable than reducing kurtosis or percentages of zero values. Therefore, when high kurtosis, a larger percentage of zero values, and skewed data are present in transaction data, applying both a Softmax transformation as well as the Doughnut clustering method is more advisable.

In our current data generation procedure, we only assume non-correlated data. How effective the Doughnut clustering method works on correlated transaction data is still an open question. As a result, taking into account correlation between different product variables serves as an interesting future study direction. In addition, the ratio of between-cluster variance to within-cluster variance is often used as validity metric for hierarchical clustering algorithms in the literature. Since the Doughnut clustering method is based on the non-hierarchical algorithm K-means, it may also be interesting to explore other validity metrics in addition to RSQ_ratio in our future research.

REFERENCES

Baer, Darius. (2012). *CSI: Customer Segmentation Intelligence for Increasing Profits.* Paper presented at the Proceedings of the SAS Global Forum 2012 Conference, Cary, NC.

Baer, Darius, & Chakraborty, Goutam. (2013). *Product Affinity Segmentation Using the Doughnut Clustering Approach.* Paper presented at the Proceedings of the SAS Global Forum 2013 Conference, Cary, NC.

Collica, Randall S. (2011). *Customer Segmentation and Clustering using SAS® Enterprise Miner™ Second Edition*. Cary, NC: SAS Institute, Inc.

Fleishman, AllenI. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*(4), 521-532. doi: 10.1007/bf02293811

Halkidi, Maria, Batistakis, Yannis, & Vazirgiannis, Michalis. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems, 17*(2-3), 107-145.

Jain, Anil K., & Dubes, Richard C. (1988). *Algorithms for clustering data*: Prentice-Hall, Inc.

Reinartz, Werner J., Echambadi, Raj, & Chin, Wynne W. (2002). Generating Non-normal Data for Simulation of Structural Equation Models Using Mattson's Method. *Multivariate Behavioral Research, 37*(2), 227-244. doi: 10.1207/S15327906MBR3702_03

Rousseeuw, Peter J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics, 20*, 53-65.

Vale, C. David, & Maurelli, VincentA. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48*(3), 465-471. doi: 10.1007/BF02293687

Wicklin, Rick. (2013). *Simulating Data with SAS*: SAS Institute.

CONTACT INFORMATION

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

Name: Juan Ma
Enterprise: Oklahoma State University
Address: 91 S. University Place Apt #12
City, State ZIP: Stillwater, OK 74075, USA
E-mail: juan.ma@okstate.edu

Name: Darius Baer
Enterprise: SAS Institute Inc
Address: 100 SAS Campus Drive
City, State ZIP: Cary, NC 27513, USA
E-mail: Darius.Baer@sas.com

Name: Goutam Chakraborty
Enterprise: Oklahoma State University
Address: 420 A Spears School of Business
City, State ZIP: Stillwater, OK 74078, USA
E-mail: goutam.chakraborty@okstate.edu

Juan Ma is a Ph.D. candidate in the School of Industrial Engineering and Management at Oklahoma State University (OSU). She completed the SAS and OSU Data Mining Certificate program requirements in 2012. She has earned three SAS certifications: Certified Predictive Modeler using SAS® Enterprise Miner™ 6.1, SAS Certified Base Programmer for SAS® 9, and SAS Certified Advanced Programmer for SAS® 9. She made it to the top three in the SAS data mining shootout competition twice. She is a co-recipient of the 2nd place in 2012 and 3rd place in 2014. She also presented her work in the competitions at the 2012 Analytics Conference and the 2014 Analytics Conference respectively.

Dr. Darius Baer is an advisory analytical consultant at SAS Institute. He has over 34 years of SAS® experience using statistical methods to solve executive driven business problems for a variety of

industries including retail, pharmaceuticals, manufacturing, telecommunications, finance, government, and others. He is the developer of the Doughnut Clustering Method which has been successfully implemented in a wide range of businesses. He has also developed other techniques to improve analytic predictions including the Anomaly Detection and Cardinality Reduction Methods.

Dr. Goutam Chakraborty is Ralph A. and Peggy A. Brenneman professor of marketing and founder of SAS® and OSU data mining certificate and SAS® and OSU marketing analytics certificate 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 has over 25 years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.