Simultaneous Co-Clustering and Classification in Customers Insight

M Anggistia , A Saefuddin, and B Sartono

Department of Statistics, Bogor Agricultural University, Indonesia

Corresponding author: meta.anggist@gmail.com

Abstract. Building predictive model based on the heterogeneous dataset may yield many problems, such as less precise in parameter and prediction accuracy. Such problem can be solved by segmenting the data into relatively homogeneous groups and then build a predictive model for each cluster. The advantage of using this strategy usually gives result in simpler models, more interpretable, and more actionable without any loss in accuracy and reliability. This work concerns on marketing data set which recorded a customer behaviour across products. There are some variables describing customer and product as attributes. The basic idea of this approach is to combine co-clustering and classification simultaneously. The objective of this research is to analyse the customer across product characteristics, so the marketing strategy implemented precisely.

1. Introduction

Competition in current business is tighter than a few years ago. High profit always became the priority indicator for the success of the company. Campaign product is important for customers to buy products. However general campaign model usually does not fit particular customer's interest. Customer diversity may be addressed by the different type of campaign strategy [1].

Market segmentation is one of the popular marketing strategies to face customer diversity. Smith in [2]were mentioned that "Market segmentation is to divide a market into smaller groups of a buyer with distinct needs, characteristics, or behaviours who might require separate products or marketing mixes". In other words, segmentation is a strategy to divide customers become homogenous groups. Sun [2] explained six roles of market segmentation as the following: 1) to choose customers interest and desire appropriately, 2) to seek market advantages, 3) to determine features in market aim, 4) to make strategy for marketing mixes, 5) helping small company to develop business and survive among another group, and 6) easier to get feedback and arrange marketing strategy. Based on these phenomena, statistical analysis plays a crucial role. Study of marketing segmentation classified in two, these are supervised and unsupervised technique. Clustering is one of the unsupervised methods. According to [3], clustering may reduce the data dimension by forming groups with a particular characteristic in each cluster.

Clustering by combined with grouping form two ways is called co-clustering. Cheng [4] were described co-clustering as simultaneous column and row grouping from matrix data set. Each co-cluster will have homogeneous observations. Co-clustering is more favourable than clustering in some viewpoints, i.e. more informative, use fewer parameters, and time reduces in computation [5]. The objective of the research is to apply simultaneously co-clustering and logistic regression approach in transaction data set. So that analysing influence factors in each co-cluster can be done by logistic regression. Therefore offered campaign is more efficient because attained by specific targeted customers.

2. Related Work

One of the popular marketing strategies is segmentation, targeting, and positioning (STP). Segmentation is grouping the homogenous clients so that each cluster will have distinct needs and desire while targeting is direct marketing activities at each cluster which customers believe that the company can satisfy better the competitors. Moreover, positioning is a position the product to offering the customers to buy. Therefore, making the segment is important in the marketing strategy [6]. In [2], it mentioned that there are four bases in segmentation, i.e. geographic, psychographic, demographic, and behaviour segmentation. Among all segmentations, demographic is the most popular one. Partly because of the customer desire, preference, and usage rate are related to demographic variables. Segmentation in statistical analysis can be done by using supervised or unsupervised learning. The supervised learning uses the known corresponding target training data set to give feedback after the procedure finished. The supervised learning involves building a regression or a classification model. The unsupervised learning is data categorization without any corresponding target values. The unattended is grouping the data by finding the correlation from the raw dataset, where it is called clustering. Singh [7] was used the clustering method to segmenting the customer using RFM (recency, frequency, and monetary) variable from transaction data set. The group of clients, which have high RFM value will be chosen as a marketing target because they give great profit for the company while the small one needs another market strategy.

Haider [8] were improved the segmentation technique using clustering and classification method simultaneously. This method divides into two main steps. The first step is clustering the observation, while the second phase is model building. The objective is put the new customer to the proper group. The algorithm uses iteration to find the best members of the cluster, but it does not guarantee the objective function will increase monotonically. So the highest objective value is the best choice.

Clustering technique improved from years to years. The researchers want to group not only from one way but also from two ways. In 2000, Cheng [4] were proposed co-clustering using expression data set. The expression dataset builds from gene as a row across condition as a column. The co-clustering in an expression dataset will generate a homogeneous group of genes and conditions by minimising the mean square error. Each co-cluster will have different characteristics of genes and conditions. However, the observations inside the co-cluster will have a high similarity to the interpretation is easier. Cho [9]were developed the Cheng and Church's method. They proposed sum square error as the objective function, while Cheng and Church used mean square error. The best co-clustering determines by minimised the objective function. When the Cheng and Church's co-clustering method founds one co-cluster each time, the Cho *et al*.'s method could find $k \times l$ co-clustering each time. So this new method is more efficient in computational time. Co-clustering can be written down as a mapping from *m* rows data to *k* rows cluster that symbolise with ρ and mapping from *n* columns data to *l* columns group that express with γ. The mapping is written down by:

$$\begin{aligned} \rho &: \{1,2,3,\ldots,\mathrm{m}\} \rightarrow \{1,2,3,\ldots,\mathrm{k}\} \\ \gamma &: \{1,2,3,\ldots,\mathrm{n}\} \rightarrow \{1,2,3,\ldots,\mathrm{l}\} \end{aligned} \tag{1}$$

With $\rho(i) = g$ interpreted as a row i is assigned to row cluster g and $\gamma(j) = h$ construed as the column j assigned to the column grou h p. Cho [9] used iteration to find the best objective value. The repetition guarantees the objective function monotonically decreased. So, compared with the previous method, this method is more efficient in computational time.

The co-clustering method also can be used in text mining problem. Dhillon [3] were did the co-clustering using the data of words across documents. This approach uses information theoretic as the objective function. The aim in the previous co-clustering method is to minimising the objective function, but this research's aim is to maximising the objective function. The iterations in this approach ensure the objective function monotonically increased. The result from this method concludes that a high dimension and heterogeneous data handled. Another co-clustering technique is collaborative filtering, which proposed by George [10]. By using the movie rating dataset, this method can predict the missing data. The missing data from each co-cluster predicts with the average from non-missing data in their co-cluster. The better method of missing data prediction in co-clustering was developed by Banerjee [5]. This research uses the Bregman divergence, and it gives a better result than collaborative filtering in predict the missing data.

In general, the supervised approach captures global structure but disregards any local structure in response measurement. On the other hands, the unsupervised approach focuses exclusively on achieving a local structure in response measurement. Therefore Agarwal [11] were proposed predictive discrete latent factor models. The idea combines the benefits of both supervised and unsupervised learning approach. Expected-Maximization(EM) algorithm used in this method. The E-step does renewal the members in each co-cluster while M-step does renewal the prior function in each co-cluster. The dyadic data structure is the data with response value that can describe in two (or more) interconnected group at the matching point. This data consists of two data matrix. The first data is an array of response value. While the second data is a set of explanatory variables called covariate vector. The covariate vector forms from row attributes, column attributes, and another particular feature of column and row pair. In 2010, Dheodar [12] were introduced another co-clustering algorithm using the dyadic data structure as a covariate. The method is simultaneous co-clustering and classification. The algorithm explained in the next chapter. This algorithm applied in this research by using transaction data set from cosmetic and body care company.

3. Simultaneous Co-Clustering and Classification

The response variable **Z** is sized $m \times n$. With *m* is the number of rows and *n* is the number of columns. The value of reply deals with a binary class, which coded -1 if a customer buys a product and 1 if a customer does not purchase a product. The covariate vector $\mathbf{x_{ij}}$ forms from row attribute $\mathbf{c_i^T}$, column attributes $\mathbf{p_j^T}$, and another particular feature of row and column pair $\mathbf{a_{ij}^T}$. Logistic regression is used to model building in classification method. Because response values are binary, the transformation needed in logistic regression. The conversion makes the model into linear. The change is called logic. The linear equation for a linear model written by:

$$\ln\left[\frac{P\left(z_{ij}=1 \mid x_{ij}\right)}{1-P\left(z_{ij}=1 \mid x_{ij}\right)}=\beta^{T} x_{ij}\right] \tag{2}$$

with $\boldsymbol{x_{ij}}^{T}=\left[1, \boldsymbol{c_i^T}, \boldsymbol{p_i^T}, \boldsymbol{a_{ij}^T}\right]$ and $\boldsymbol{\beta^T}=\left[\boldsymbol{\beta}_0, \boldsymbol{\beta_c^T}, \boldsymbol{\beta_p^T}, \boldsymbol{\beta_a^T}\right]$.

The aim of this research is simultaneously clustering row and column. Which are called co-clustering? So will be formed *k* row clusters and *l* column clusters or $k \times l$ co-cluster (block). Each co-cluster will have different model and characteristic. To get the best co-cluster (ρ,γ), minimising the loss function is needed. The loss function created from both co-clustering and logistic regression. The loss function defined as

$$\sum_{g=1}^{k} \sum_{h=1}^{l} \sum_{u:\rho(u)=g} \sum_{v:\gamma(v)=h} \ln\left(1+\exp\left(-z_{uv} \beta^{gh^{T}} x_{uv}\right)\right) \tag{3}$$

With $\mathbf{z_{uv}}$ is the response value in a row *u* and column *v,* and $\boldsymbol{\beta^{gh}}$ is the vector coefficient in each model from sub-matrix **Z**. In simultaneous co-clustering and classification, initialization for co-clustering (ρ,γ) in matrix **Z** takes first. After initialization, the next is model building using logistic regression that is changing the row in the best row cluster by calculating row error. Row error is formulated by

$$E_u(g)=\sum_{h=1}^{l} \sum_{v:\gamma(v)=h} \ln\left(1+\exp\left(-z_{uv} \beta^{gh^{T}} x_{uv}\right)\right) \tag{4}$$

The row goes into a row cluster if the row error value in that row group is the smallest one. In other words row, *u* is added in row group *g* if $\rho^{baru}(u)=\arg\min_g E_u(g)$. After all the rows added in the best row cluster, the next changes the column to the best column group by using the same step in the row changing. The column error can be defined by

$$E_v(h)=\sum_{g=1}^{k} \sum_{u:\rho(u)=g} \ln\left(1+\exp\left(-z_{uv} \beta^{gh^{T}} x_{uv}\right)\right) \tag{5}$$

Input: $Z_{m \times n}$, covariate
Output: Co-clustering (ρ,γ) and co-clustering model βs

1. Begin with a random co-clustering (ρ,γ)
2. Repeat

Step 1

3. Update co-cluster models
4. For g=1 to k do
5. For h=1 to l do
6. Build logistic regression model each co-cluster (g,h), to obtain update β^{gh}
7. End for
8. End for

Step 2a

9. Update ρ – assign each row to the row cluster that minimises the row error
10. For u=1 to m do

$$\rho(u) = \arg\min_g \sum_{h=1}^{l} \sum_{v:\gamma(v)=h} \ln\left(1+\exp\left(-z_{uv}\beta^{gh^T}x_{uv}\right)\right)$$

11. End for

Step 2b

12. Update γ – assign each column to the column cluster that minimises the row error
13. For v=1 to n do

$$\gamma(v) = \arg\min_h \sum_{g=1}^{k} \sum_{u:\rho(u)=g} \ln\left(1+\exp\left(-z_{uv}\beta^{gh^T}x_{uv}\right)\right)$$

14. End for
15. Optional: Repeat 2a and 2b until convergence

Until convergence

16. Print (ρ,γ) and β

Source: Dheodar [12]

Figure 1. Algorithm of simultaneous co-clustering and classification

Column moving will make the best row cluster arrangement different. Likewise, row moving will make the best column group method different. Therefore iteration is needed until the column and row assignment not changing. The sign of that is the convergence of loss function. Finding the best co-cluster does not mean that finding the best model. So the iteration is also needed in the parameter estimation and co-clustering. The iteration held until the loss function is convergence in one value. Note that iteration ensures the loss function is monotonically decrease. But the iteration in this approach just convergence in a local minimum. A simple algorithm was proposed by Dheodar [12]. As can be seen in Figure 1 there is two first step of this method. The first step is minimising objective function using logistic regression, while the second phase is exchanging row and column to obtain optimal co-cluster. The algorithm needs some initialization on co-cluster membership which and then optimise during the iteration. As often most optimisation algorithm, various initialization may produce a different final result. So the authors suggest to runs the algorithm several times and select the best result with smallest loss value.

Working with a huge number of observations may lead very long computation time in estimating logistic regression model. Instead of using the complete set of observation, we propose to modify the algorithm by using a reasonably small subset of view. In this paper, we randomly select 10% remarks from each co-cluster to obtain the logistic model. The reader may use the other portion depend on the size of the data set.

4. Experimental Results

4.1. Data Description

A beauty and body care company produce many types of products, i.e. makeup, body care, and perfume. A lot of number branches in the different point of sale area make the diverse customer characteristics can't be avoided to facilitate the marketing product, and the company was applied the membership card for the regular customer. Fortunately, 90% of the customers have a membership card. Therefore the company can store the client data such as age, address, marital status, gender, and so on.

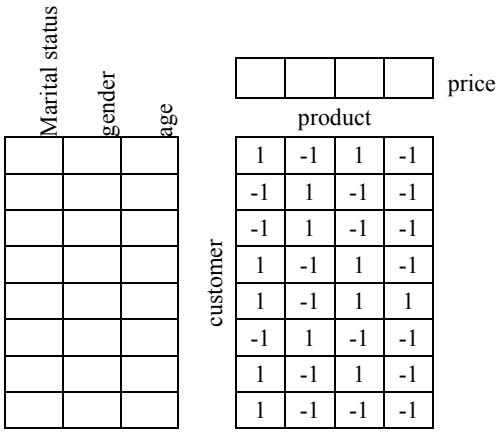


Figure 2. Data illustration

The various customers make the company do the segmentation the customer. Along with many of products, segmentation of customer across product is a need. The client segment information used for making marketing strategy, such as campaign held in the right place for the right customer while products segment used for product development like pricing strategy. The company always records the product transaction. The documents are called transaction data set, will be employed in this paper. The transaction data set records 198,537 members who purchased the product in four months. Furthermore, 66 selected products used in this analysis. The data of purchased product by the customers presented in matrix **Z** with the binary value. The value 1 in matrix **Z** means the customer buys the product and -1 means the customer did not purchase the product.

This paper was not only using transaction as primary data set but also the auxiliary data set. The auxiliary consists of the customer attribute and the product attribute. The variables that can explain the characteristics of the client are age, gender, and marital status. That three variables were called customer attributes $\mathbf{c}_i$ while the product quality $\mathbf{p}_j$ is just a price.

On the data analysis, the customers denoted as row and products as a column, so 198,537 × 66 is the size of matrix **Z**. The customer attributes called as matrix **C** has size 198,537 × 3 and product quality will be matrix P with size in 66 × 1. Figure 2 displays the data illustration. The number of customer clusters and product clusters was objectively chosen as 5 and 4 by authors. As mentioned above, running data analysis in several times can't be avoided, so this study decides to use 5 repeats. The result from the repetition is the fourth one has smallest loss value. So the fourth repetition is chosen to clustering and logistic regression for this paper. Figure 3 display the line plot of objective value in the fourth repetition. By looking at Figure 3, there are 11 iterations of this algorithm to makes the real value convergence. The smallest loss value is 1,983,035.

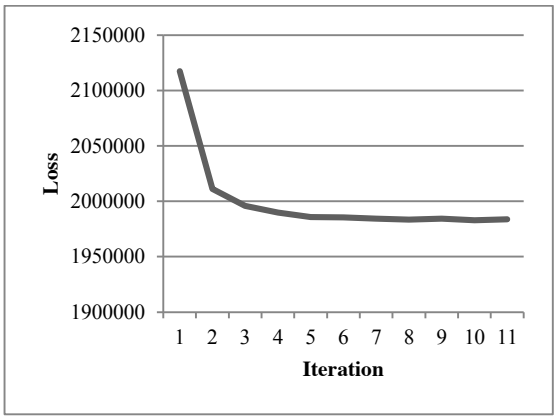
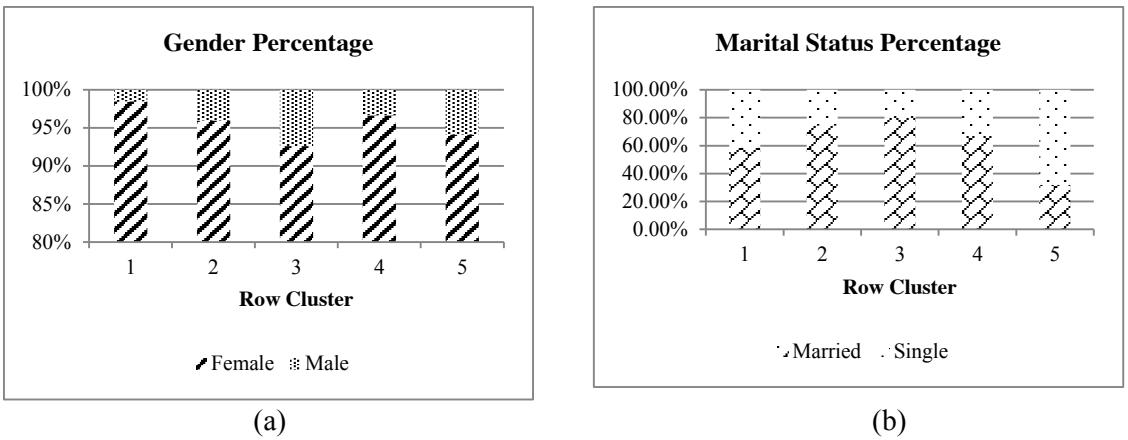


Figure 3. Line plot of loss value in 4th replication

4.2. *Result*

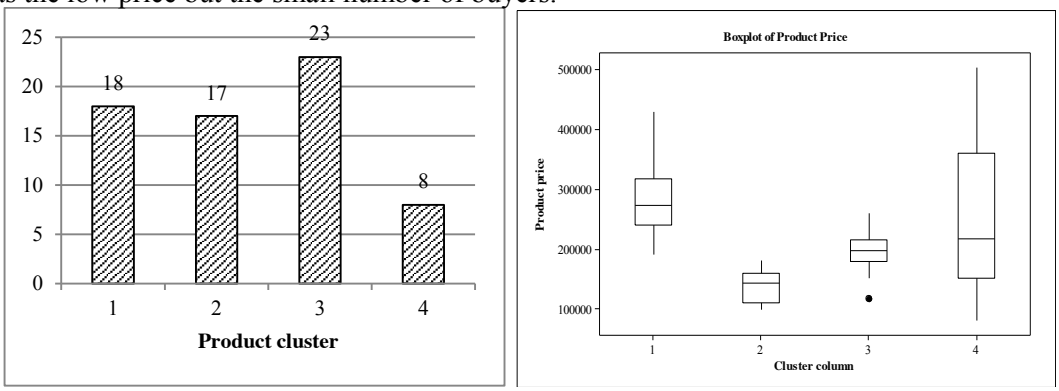
In each customer cluster, the number of females always has a higher percentage than male. This result fits in this problem because the data taken from a transaction of cosmetic and body care company. Furthermore, the customer cluster number three and five have more male customers than the other cluster. However, in customer number three, the married customers are dominated, but the single customer in customer group number five is dominating. That means young clients, both female and male were gather in customer group number 5 and the old ones were gather in customer cluster number 3. Product group number three collects 23 products with slightly low price than others. There is an outlier in product category number three, which the outlier has the lowest price for the other product. The product group number five only has 8 products, but the box plot shows that the product's price is diverse and has positive skewness. This skewness concluded that there are products in this cluster that have a higher price than the average price. Same with the product price in the product group number 4, the product cluster number 1 is also gathered products with the high price.



(a) (b)

Figure 4. Row cluster characteristics (a) Number of gender each row group in percentage (b) Number of marital status each row cluster in percentage

The percentage of buyers in each co-cluster displayed in Table 1 and the plot of customers across consumer show in Figure 6. The black dot means the customer purchased the product. Product cluster number three almost has full of dots. So in product group number three, the products are the favourite one. The price in product category three is low enough maybe this is one of the reasons. But the cheap one does not mean that the product will takes customer interest because the product cluster number two has the low price but the small number of buyers.



(a) (b)

Figure 5. Product group characteristic (a) Number of product (b) Box plot of product price

The second favourite products gather in the product category number four. The products price is quite high than the products cluster number three. So, it concluded that the product group number four gives superior profit for a company. Otherwise, the product category number one and two have a small number of buyers. It is shown in Figure 6 that the black dot just a little. The product cluster with a low percentage of consumers indicates that the goods do not catch the customer interest, so this group of the goods need some development.

Table 1. Percentage of buyers in each co-cluster

Customer cluster	Product cluster			
	1	2	3	4
5	1.42%	1.09%	9.18%	9.48%
4	0.25%	1.59%	8.88%	4.57%
3	1.37%	1.79%	13.58%	9.78%
2	1.68%	0.22%	3.67%	12.37%
1	0.02%	0.69%	4.26%	1.54%

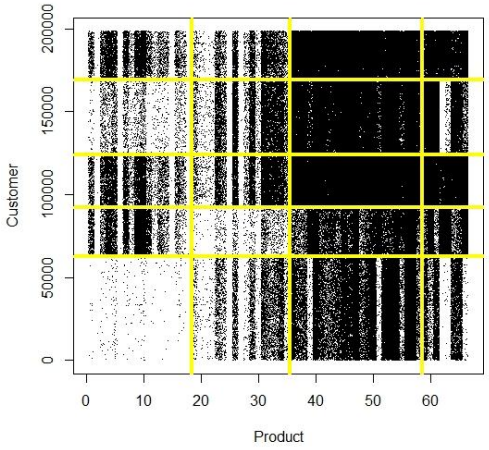


Figure 6. Plot of the product purchase by customer

The customers can be chosen as a market target if they like to purchases many products. That the clients are on the co-cluster number (3,3), (2,4), (5,4), (5,3) and (4,3). The other side, the co-cluster with a small number of buyers such as co-cluster number (1,1), (2,2), (4,1), (5,1), and (1,4) need a better marketing strategy to increase the number of buyer and frequency of products purchase. Product cluster number one and two show that there are around 5 white lines. The White line in Figure 6 means that the goods are rarely bought by the customer. The company should take action from this case. The better market strategy can save the product with catch the client's interest to buy the products. Figure 6 shows in co-cluster (4,4) some products rarely purchased. While in co-cluster (3,4) and (5,4) that products were many purchased by customers. So, it concluded that the clients in co-cluster (3,4) and (5,4) interest in this products, while the customers in co-cluster (4,4) not take interest for this products. This products campaign for the clients in co-cluster (3,4) and (5,4) may lead to a high frequency of purchasing the products. The coefficient of logistic model for each co-cluster displayed in Table 2. All intercept coefficients are significant in alpha 5%; this indicates that there are another variables that would make the customer purchase the product. Beside intercept factor, almost all the price coefficients are significant in all models, so it concluded the product price changing gives influence for the customer decision to buy a product. If the price decreases then the probability of the client buys product higher because the cost coefficient has a negative sign. The high value of fixed price ratio means the customers are sensitive in price changing.

Table 2. Logistic regression coefficient

Customer cluster	Variable	Product cluster			
		1	2	3	4
1	Intercept	0.62*	-1.53*	-1.75*	-3.16*
	#Customer				
	Gender	0.04	0.17	0.14*	0.02
	Age	0.00*	0	0.00*	0.00*
	Marital status	-0.03*	0.04	0.04*	-0.01
	#Product				
	Price	-12.87*	-21.16*	0.97*	-0.17
2	Intercept	-0.93*	-4.34*	-2.28*	-7.38*

	#Customer				
	Gender	-0.10*	-0.48	0.02	-0.11
	Age	0	0.01	0.00*	0.01*
	Marital status	-0.09*	-0.33	0.05*	-0.12*
	#Product				
	Price	-13.21*	-31.77*	-3.09*	8.03*
3	Intercept	2.49*	-1.36*	-1.15*	-0.37*
	#Customer				
	Gender	-0.07*	0.21*	0.60*	0.03
	Age	-0.01*	0.01*	0.01*	0.01*
	Marital status	-0.02	-0.04	0.11*	-0.02
	#Product				
	Price	-28.01*	-34.33*	-20.05*	-24.29*
4	Intercept	-3.78*	-5.58*	-5.37*	-8.92*
	#Customer				
	Gender	0.48*	1	0.48*	-0.13*
	Age	0.02*	0.01	0.01*	0
	Marital status	-0.12*	-0.19	0.04*	0.02
	#Product				
	Price	-3.02*	-23.8	8.59*	20.86*
5	Intercept	0.83*	-1.14*	-1.30*	-2.92*
	#Customer				
	Gender	0.03	0.05	0.13*	-0.16*
	Age	0.00*	0	0.00*	0.01*
	Marital status	-0.07*	-0.08	0.05*	-0.07*
	#Product				
	Price	-16.83*	-30.98*	-4.66*	-4.81*

*significant in alpha 5%

All marital status coefficient in product cluster number three has a positive value and great. That indicated that the marital status gives influence for customer decision, such as the probability of a married customer to buy the product in product cluster three is higher than a single client. Not significantly of the gender coefficient in co-cluster means that the goods usually purchase both female and male customer. Likes in the co-cluster (1,1), (1,2), and (1,4), the probability of the product purchase for male and female is same. That supported by the percentage of buyer that displays in Table 1. The not significant of age coefficient means the product can use for all ages.

5. Summary

Simultaneous co-clustering and classification is the right method for grouping observations and building models in a dyadic data structure with a covariate. Note that in this research assume that there are no missing data. If there are missing data, weight included in the loss function calculation. The repetition in this approach is needed because initialization takes place in the first step. So the number of repetitions may lead to the better result. While the huge number of observation will lead a long computational time, sampling used as the one of a solution.

Based on the result, the simultaneous co-clustering and classification give the significant result in transaction data set with covariates. The product cluster with small numbers of buyer needs product improvement or attracts the customer interest for that product. The products with high consumer interest gather in the client cluster number three. Not high product price may be one of the reasons the product became popular. For the company, use at least one-year transaction data set may give a better result for the customer and product segmentation than use four months data set because each month has a different event.

References

[1] Goyat S 2011 EJBM **3** 45 (http://www.iiste.org/Journals/index.php/EJBM/article/view/647)
[2] Sun S 2009 *IJBM* **4** 63 doi: 10.5539/ijbm.v4n2p63.

[3] Dhillon I, Mallela S and Modha D 2003 *Proc. of the Ninth ACM SIGKDD Int. Conf. on Know. Disc. and Data Min* p 89 doi:10.1145/956750.956764.
[4] Cheng Y and Church G 2000 *Proc. Int. Conf. Intell. Syst. Mol. Bio.* (*ICMB*) **8** *p* 93 (https://www.ncbi.nlm.nih.gov/pubmed/10977070)
[5] Banerjee A, Dhillon I, Ghosh J, Merugu S and Modha D S 2007 *JMLR* **8** 1919 (http://www.jmlr.org/papers/v8/banerjee07a.html)
[6] Lynn M 2011 *Segmenting and targeting your market: Strategies and limitations* (In Sturman M C, Corgel J B and Verma R (Eds.)) The Cornell School of Hotel Administration on Hospitality: Cutting edge thinking and practice (pp. 353-369). Hoboken, NJ: Wiley. (http://scholarship.sha.cornell.edu/articles/243)
[7] Singh A, Rumantir G, South A and Bethwaite B 2014 *Proc. of the 2014 Int. Conf. on Big Data Sci. and Comput.* doi: 10.1145/2640087.2644161.
[8] Haider P, Chiarandini L and Brefeld U 2012 *Proc. of the 18th ACM SIGKDD Int. Conf. on Know. Disc. and Data Min. p* 417 doi: 10.1145/2339530.2339600.
[9] Cho H, Dhillon I, Guan Y and Sra S 2004 *SIAM Int. Conf. on Data Min* (SDM) *p* 114 (http://bigdata.ices.utexas.edu/publication/minimum-sum-squared-residue-co-clustering-of-gene-expression-data)
[10] George T and Merugu S 2005 *Proc. of the Fifth IEEE Int. Conf. on Data Min.* (*ICDM*) *p* 625 doi: 10.1109/ICDM.2005.14.
[11] Agarwal D, Merugu S. 2007. *Proc. of the 13th ACM SIGKDD Int. Conf. on Know. Disc. and Data Min. p* 26 doi: 10.1145/1281192.1281199.
[12] Dheodar M and Ghosh J 2010 ACM *Transactions on Knowledge Discovery from Data* (*ACM TKDD*) **4** doi:10.1145/1839490.1839492.