Sample Identification Approach by K-Means Clustering In Thinner Retail Market Segmentation

Langgeng Listiyoko[1*], and Marhaendro Purno[2]

[1] STMIK Muhammadiyah Banten, Tangerang, Indonesia
[2] Management Department, Sekolah Tinggi Ilmu Ekonomi Insan Pembangunan, Tangerang, Indonesia

*langgeng.listiyoko@stmikmbanten.ac.id

Abstract. Ingredients identification of thinner sample in retail market is very easy to do by a special machine, and then the product would be copied. Then the problem is how to set the sample of competitor product into the segmentation based on many consideration aspects. Data mining helps user to identify whether the sample is a member of one segmentation or not based on the closest characteristic value that observed. K-means clustering calculate a numeric value of each sample product's characteristic then classify into a number desired cluster. Data history has 21 existing products and classified into 4 cluster at the beginning, then two data tests (competitor products) put into the data set to identify what is the nearest cluster. The result of K-means clustering shows the first competitor as cluster_1 while the second one is cluster_3.

1. Introduction

Organizations are rapidly investing in developing strategies for better customer acquisition, maintenance and development. The concept of business intelligence has a crucial role to play in making it possible for organizations to use technical expertise for acquiring better customer insight for outreach programs [1]. In the retail business, common way to fight competitor is by doing sample analysis in the research and development laboratory. There is a typical problem in product development laboratory related to the competitor product identifying. That is not only about what their content in the product but also market segmentation as well. It is hard to maintain every single new product triggered by new sample competitor regarding number of raw materials, formulation, over volume stock, etc. New product launching must be decided wisely identify market segmentation first to see the possibility compare with existing product first. A simple to do is comparing all of existing product with a certain sample of competitor. However, this activity needs much more time to complete the analysis. If the sample identified clearly into the segmentation, then analysis activity will be focused on one segment that membered by a several existing products.

Market segmentation is one of the most fundamental strategic marketing concepts [2]. One of segmentation technique is cluster analysis, which is a term that refers to a large number of techniques for grouping respondents based on similarity or dissimilarity between each other. The similarity could be construct by multi variable that describes the complexity of item attributes.

Cluster analysis has been employed in the development of potential new product opportunities. By clustering brands/products, competitive sets within the larger market structure can be determined. Thus, a firm can examine its current offerings vis-a-vis those of its competitors [3].

Cluster analysis for creating market segmentation is common activity in industries and getting more popular day by day. It has become a common tool for the market researcher [3]. The character as result of cluster analysis considered in decision making regarding item product to be launched and or maintained. In the paint industries competition, that is produce many types of thinner it is important to maintain the product line so as the raw material simplified. The ingredients of thinner can be extracted easily by gas chromatography extractor machine, means that the formula not quite secure anymore. The idea of this research is comparing the competitor product with the existing one after the ingredients identified to see about the nearest character in certain cluster that created before.

2. K-Means Clustering

There is a different approach in cluster analysis that realized by computer science and marketing research. In the simple marketing research, cluster analysis can be constructed by Customer Relationship Management (CRM) to figures what the customer need by evaluate their spending and purchasing habits [1]. It seems no complexity in the project, while computer science observes more detail about the customer attributes to find out and build an accurate pattern. Instead, accurate clusters can be obtained using prototypes of similar time series [4]. K-Means clustering algorithm is an unsupervised classification process [5] of a data set. It uses the historical data that prepared to identify and build the pattern, and then implemented to whenever trained data added. Some author combines k-means with another available algorithms such as AHP (Analytical Hierarchy Process) related to the decision support system, or just by sorting either ascending or descending techniques. *K*-means is not a hierarchical clustering algorithm, but a relocation method [6][7]. It splits data set into a number of expected clusters, do a calculation distance to the centroid, and relocated a member of cluster when new calculation of each data get closer than the existing member.

Talking about Knowledge Discovery in Database, there are 9 steps to get comprehensive pattern [8]: 1) Domain Understanding and KDD Goal, 2) Selection and Addition, 3) Pre-processing Data, 4) Transformation, 5) Choose the appropriate Data Mining Task, 6) Choose Data Mining Algorithm, 7) Employing Data Mining Algorithm, 8) Evaluating and Interpreting, 9) Visualization and Integration.

While CRISP-DM (Cross Industry Standard Process for Data Mining) is more suitable [9] since CRISP provides a non-proprietary and freely available standard process for fitting data mining into the general problem-solving strategy of a business or research unit, while user of this research is product development department. CRISP-DM consists of six phases, illustrated in figure 1.

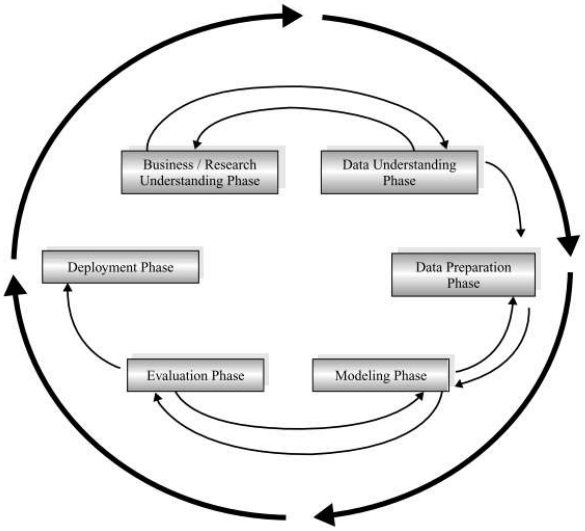

Figure 1. CRISP-DM Phases [9]

2.1. Business / Research Understanding Phase
In this phase, project objectives and requirements are clearly determined, translate the goal and restriction into the DM formulation, and prepare the preliminary strategy for achieving the objectives. The goal has to be understood by all the team member.

2.2. Data Understanding Phase
This phase is related to all of data attributes. Start with data collecting, even they come from any other different sources, and or format. An initially exploratory data and analyze is needed to discover initial insights. Evaluate quality of data, and selecting the subset that may contain any actionable pattern then completing the data understanding phase.

2.3. Data Preparation Phase
Data preparation is completely labor intensive while data has to be prepared about initial data until the final that is to be used for all subsequent phase. It is also about selecting the appropriate variables (since only related variable will be used in the modelling), transforming data, and cleaning data as well. Data cleaning means only use the complete attribute value of each data record, no null data needed.

2.4. Modelling Phase
Determining the appropriate modelling techniques in this phase is important. It drives the researcher to prepare raw data and bring it into the line with the technique is used. So that why eventually loop process into the data preparation is needed. Even the problem can be solved by any other different data mining techniques, calibrate the modelling setting in this phase helps to optimize results.

2.5. Evaluation Phase
Evaluate the model from modelling phase regarding the quality and effectiveness before deployment, determine the best model for the research clearly must be done in this phase. The result quality depends on the raw data quality and how effective data mining modelling used.

2.6. Deployment Phase
After all the phases, deployment of modelling selected can be implemented. In this research, data mining k-means technique will be implemented. Furthermore, it will be compared with the conservative thinking as the state of the art in product development.

3. Result and Discussion
According to the CRISP-DM phases, it is important to determine and understand what the research for. The research is attempt to identify competitor product in the existing product line, considering raw material stock, component simplification regarding a simple maintenance. The main goal is how to compete in the market segmentation by existing product, so the production volume will increase, followed by cheaper raw material since company has to purchase more from the suppliers.

Raw data supplied is the existing product design, represented by formulation and competitor product content. For the confidentiality reason, each component formulation in this research will be encrypted by a certain unique code. Here Table 1 is initial data product component to be processed.

Table 1. Initial raw data

	S0003	S0002	S0004	S0006	S0007	S0008	S0009	S0010	S0011	S0012	S0013	S0014	S0015	S0016	S0018
T0022		43						6,5			2	5		43	
T0024		20		30				41,25			8,75				
T0010		15												70	
T0025		20						51,6			28,4				
T0023					35	29		24,5			11,5				
T0017													96,6		3,4
T0016			52,5					9,1	19						19,4
T0013			57,8					14,9							
T0014					35	30	10		15						
T0020	25		50												
T0021			11			46		6	21	4				4	
T0019									50						50
T0001								36	10						34
T0002							34	31	15						15
T0003						20		20	15						35
T0004					10	10	40	10	15						
T0005					41				14				26		
T0006					25				10				20		
T0007				40	35							15			
T0008					35	30	10		15						
T0009					35	15			5						35

Raw data need to be pre-processed since there are null value that could not be processed, so that need a transformation. It is so simple by doing zero replacement in the null value so the modelling can read what does it means. Data preparation including data cleaning, data cleansing that is drop out all the inappropriate record, for instance item with only come by one raw material (100% ingredient by single raw material) must be dropped since it will very easy to identify. Another case is if a raw material only used by one item product while the concentration is very low, let say under 5%, then it will be dropped too. Table 2 illustrates final condition of pre-processed data.

Table 2. Pre-processed data history

	S0003	S0002	S0004	S0006	S0007	S0008	S0009	S0010	S0011	S0012	S0013	S0014	S0015	S0016	S0018
T0022	0	43	0	0	0	0	0	6,5	0	0	2	5	0	43	0
T0024	0	20	0	30	0	0	0	41,25	0	0	8,75	0	0	0	0
T0010	0	15	0	0	0	0	0	0	0	0	0	0	0	70	0
T0025	0	20	0	0	0	0	0	51,6	0	0	28,4	0	0	0	0
T0023	0	0	0	0	35	29	0	24,5	0	0	11,5	0	0	0	0
T0017	0	0	0	0	0	0	0	0	0	0	0	0	96,6	0	3,4
T0016	0	0	52,5	0	0	0	0	9,1	19	0	0	0	0	0	19,4
T0013	0	0	57,8	0	0	0	0	14,9	0	0	0	0	0	0	0
T0014	0	0	0	0	35	30	10	0	15	0	0	0	0	0	0
T0020	25	0	50	0	0	0	0	0	0	0	0	0	0	0	0
T0021	0	0	11	0	0	46	0	6	21	4	0	0	0	4	0
T0019	0	0	0	0	0	0	0	0	50	0	0	0	0	0	50
T0001	0	0	0	0	0	0	0	36	10	0	0	0	0	0	34
T0002	0	0	0	0	0	0	34	31	15	0	0	0	0	0	15
T0003	0	0	0	0	0	20	0	20	15	0	0	0	0	0	35
T0004	0	0	0	0	10	10	40	10	15	0	0	0	0	0	0
T0005	0	0	0	0	41	0	0	0	14	0	0	0	26	0	0
T0006	0	0	0	0	25	0	0	0	10	0	0	0	20	0	0
T0007	0	0	0	40	35	0	0	0	0	0	0	15	0	0	0
T0008	0	0	0	0	35	30	10	0	15	0	0	0	0	0	0
T0009	0	0	0	0	35	15	0	0	5	0	0	0	0	0	35

Modelling selection to do this research is dedicated to K-means clustering. First, the history data need to be identified by clustering into 4 cluster. The number of cluster is totally marketing perception about how to describe the customer or market situation. Once member of each cluster is identified, new formula then will be compared the similarity so it can reduce much time of experiment because the line of laboratory experiment is targeted.

$$\boldsymbol{D(x_2 - x_1) = \|x_2 - x_1\|_2 = \sqrt{\sum_{j=1}^{P} |x_{2j} - x_{1j}|^2}} \quad (1)$$

Some literature initiated raw data given into specific cluster as it's member randomly then repeat until no more exchange of each cluster's member. Formulation (1) used to calculate distance item record value to the centroid, that is supposed as a centre of similarity in every single cluster created. Each data item compared with (in this case 4 centroid), then the membership represented by a minimum value calculation. The membership may be changing by a several iteration, until stable, no more membership changes.

In this research, initial cluster member will be identified by special software tools that are Rapid miner. Here is the output of Rapid miner, software used in the research to do clustering technique. Figure 2 describes how raw data given construct an initial cluster member, classify into 4 clusters. ID 1,2,3,4,13, and 14 are member of cluster_0 (first cluster) while next cluster has ID 5,9,11,15,16,17,18,19,20,21. Third cluster only has one member, that is ID 6, while ID 7,8,10 and 12 are member of fourth cluster.

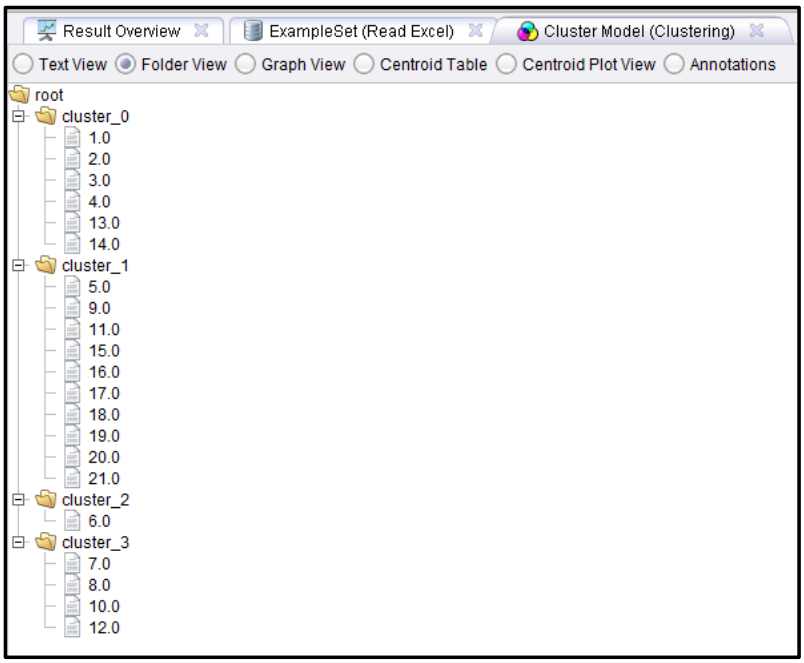

Figure 2. Rapidminer output to initiate raw data given

Model deployment then implemented to test an actual case by put new formula into the data history at once. ID 22 is new formula from pre tested competitor product by Gas Chromatography Analyser which will be identified. Figure 3 shows that the competitor product is similar to cluster_1 (second cluster), it means formulator only needs focus on the segmentation related to specification, component item, and target market as well. No more need to compare with all 21 existing product because the sample was specified to a certain segment.

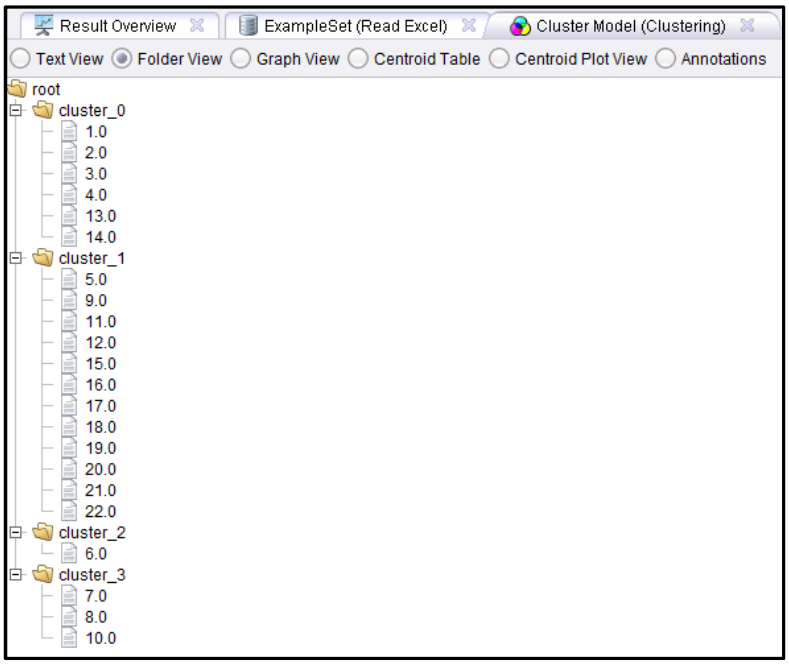

Figure 3. Rapidminer output with one data test

While figure 4 illustrates two data test that is tested by Rapidminer, put into the data history at once. We can see that data ID 1, 3 move from cluster_0 to cluster_3, followed by ID 23 that is newer data test. No changes in cluster_1 and cluster_2, it means that they are quite fix membership. We can also see that ID 22 still in cluster_2, no influenced by newer data test.

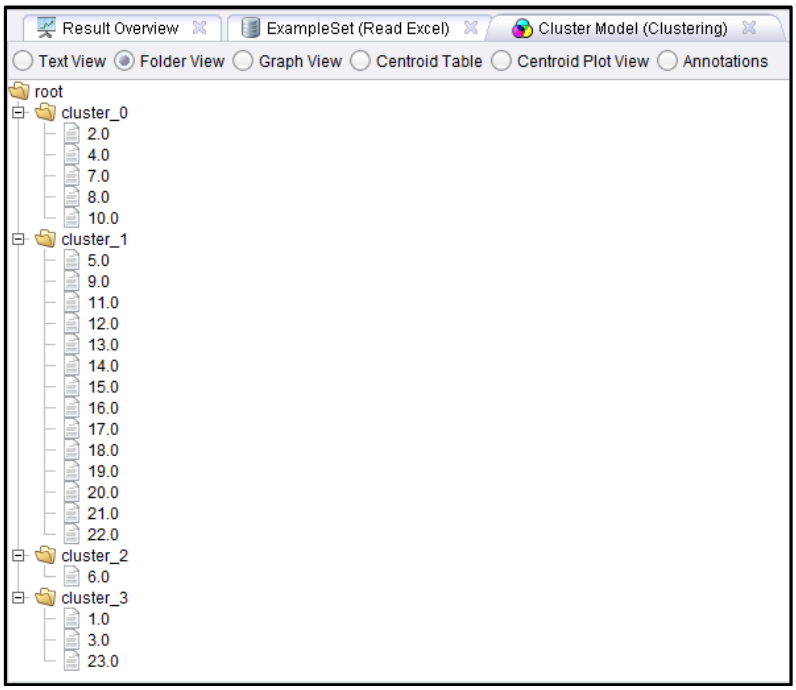

Figure 4. Rapidminer output with two data tests in time

The next step in this case after identify the data test (competitor product) is compare manually of its exactly composition. This is need to be observed deeply since every single component has a special characteristic that can affected in the finished good. Please note that K-Means and other data mining

methodologies are not expert system, so it still need man involvements. At least they save their time to do competitor product analysis since the sample will be described into a specific segment.

4. Conclusion

The first data test, that is competitor product classified into cluster_1 (second cluster), there are only 12 existing products in this segment. While the second one is classified into cluster_3, only two other existing products to be observed. Surely these results make sample analysis time become shorter. The result also helps save the time related to the laboratory experiment because cluster membership describes the similarity of each other, even they not exactly compared yet.

References

[1] S. Tripathi, A. Bhardwaj, and P. E, "Approaches to Clustering in Customer Segmentation," *Int. J. Eng. Technol.*, vol. 7, no. 3.12, p. 802, 2018.
[2] S. Dolničar, "Using cluster analysis for market segmentation - typical misconceptions, established methodological weaknesses and some recommendations for improvement," *Australas. J. Mark. Res.*, vol. 11, no. 2, pp. 5–12, 2003.
[3] G. Punj and D. W. Stewart, "Cluster Analysis in Marketing Research: Review and Suggestions for Application," *J. Mark. Res.*, vol. 20, no. 2, p. 134, 1983.
[4] S. Aghabozorgi and Y. W. Teh, "Stock market co-movement assessment using a three-phase clustering method," *Expert Syst. Appl.*, vol. 41, no. 4 PART 1, pp. 1301–1314, 2014.
[5] M. Momeni, M. Mohseni, and M. Soofi, "Clustering Stock Market Companies via K-Means Algorithm," *Kuwait Chapter Arab. J. Bus. Manag. Rev.*, vol. 4, no. 5, pp. 1–10, 2015.
[6] Oded Maimon and Lior Rokach, *Data mining and knowledge discovery*. 2012.
[7] D. Hand, H. Mannila, and P. Smyth, *Basic principles of data mining*, vol. 2001. 2001.
[8] F. Gorunescu, *Data Mining : Concepts, Models and Techniques*. Heidelberg: Springer-Verlag.
[9] D. T. Larose, *Discovering Knowledge In Data : An Introduction To Data Mining*. New Jersey: John Wiley and Sons, 2005.