J. Indones. Math. Soc.
Vol. 26, No. 02 (2020), pp. 185–191.

CLUSTERING FOR ITEM DELIVERY USING RULE-K-MEANS

Mokhammad Ridwan Yudhanegara[1,*], Sapto Wahyu Indratno[2], and RR. Kurnia Novita Sari[3]

[1,2,3]**Statistics Research Group,**
Faculty of Mathematics and Natural Sciences,
Institut Teknologi Bandung, Jl. Ganesha 10, Bandung, 40132
mridwan.yudhanegara@s.itb.ac.id, sapto@math.itb.ac.id,
kurnia@math.itb.ac.id

Abstract. In this paper, we introduce an alternative approach as model for cluster analysis. The data were analyzed by rule-k-means algorithm. It's combine between k-means algorithm and rules. As an application, we use the simulate of item delivery data to classify items based on destination addresses. The goal is to map the item based on type of delivery vehicle. The clustering can be used as a recommendation to the item delivery service company.

Keywords and Phrases: categorical feature, rules.

1. INTRODUCTION

At present, online business is very supportive of one's economy. The item delivery service industry occupies one of the central positions in the economy of modern society and is a driver of doing business both long and near. This certainly supports an increase in prosperity, especially in developed countries. Whereas in developing countries, item delivery service industry is very important to expand the development foundation and meet the increasing needs of the community.

Development of online business which is supported by the availability of item delivery services is very suitable in Indonesia, considering that Indonesia is a developing country which consists of a vast archipelago. Both the development of online bussiness and the item delivery service industry have a positive impact on human life. The positive impact of these activities is to be able to reduce the percentage of unemployment, and attract investors to invest in Indonesia.

2020 Mathematics Subject Classification: 91C20, 62H30
Received: 01-11-2018, accepted: 01-04-2019.

There are many young people and adults who start online business. Likewise accompanied by the emergence of item delivery service companies that are easily accessible in various regions. Especially for item delivery companies, item delivery activities must be carried out as effectively as possible so that the company can generate large profits. From here the company can determine how many tools and types of vehicles must be purchased, and must use public transportation modes if needed. So that online business people get good service.

In the item delivery, a lot of information is implied in the item. Based on statistical analysis, there are criteria for items that can be expressed as variables. The more criteria the item will be the more complicated the statistical analysis will be carried out. Multivariate analysis is a one of statistical method that is suitable for summarizing data with many variables.

The one of multivariate analysis that can be used to understand and simplify data interpretation is cluster analysis. Cluster analysis aims to classify objects based on the characteristics between these objects, so that they can be identified with the characteristics of each group. Specifically for item delivery services, many objects can be grouped on the item with cluster analysis such as weight, volume, accessibility of the destination address of the item, and others.

In cluster analysis, then we call clustering, is a method for finding and grouping data that has similar characteristics between one data and another data. In addition clustering is an one of the data mining methods that are unsupervised, this method is applied without training and without guidance, and does not require an output target. The data mining is a method of data processing to find hidden patterns of data, so that the results of data processing can be used to make decisions.

K-means method has been widely applied in various such as in education (Trivedi et al. [1]), general election party (Ralambondrainy [2]), credit approval and soybean desease(Huang [3]), heart desease and card credit (Huang et al. [4]), color quantization (Celebi [5]; Dhanachandra [6]), DNA Microarray (Sahu et al. [7]), etc. Next suitable for use in item delivery services. Cluster analysts that are used specifically for item delivery activities use non-hierarchical methods. Basically, there are many ways to allocate data back into each cluster during the iterative clustering process in this method. One of these methods is allocation by a strict method, where data items are expressly stated as one cluster member and not a member of another cluster. This type of method is called k-means.

Reallocation of data into each cluster in the k-means algorithm is based on a comparison of the distance between the data and centroid of each cluster. Data is allocated explicitly to clusters that have the closest data center to the data (Everitt et al. [8]; Oliveira and Pedrycz [9]; Sugiyama [10]; Anderberg [11]; MacQueen [12]). In this paper, we have the extention of k-means algorithm. The data were analyzed by rule-k-means algorithm. It's combine between k-means and rules to find new cluster.

From some descriptions of item delivery industry and information on cluster analysis, it is expected that the problem of the effectiveness of item delivery can be easily done. The goal is to map every item that will be sent to the type of item delivery vehicle. So that company owners can compete with others.

Exposure to the methods used is presented in section 2 after the introduction. For data simulation processing using rule-k-means is presented in section 3. The summary and future work of this paper are presented in section 4.

2. THE RULE-K-MEANS METHOD

The data used is item delivery simulation data that contains information item on the weight, volume, and type of road. It is must be distributed on that day. The data consist of 3 features, where 2 features are numerical and 1 feature are categorical. The categorical features needs to be quantified, here we use the weighting approach. Next, the data will be standardized.

The data processing uses data mining theory with the rule-k-means clustering algorithm. The basic k-means algorithm is given in Everitt [8], Sugiyama [10], Anderberg [11], MacQueen [12], Rencher and Christensen [13]. Next, the rule-k-means algorithm use two step.

Step 1. Let $X = \{X_1, X_2, ..., X_n\}$ be a set of n objects. $X_i = (x_{i,1}, x_{i,2}, ..., x_{i,m})$ is characterized by set of m feature. The k-means type algorithms (Anderberg [11]; MacQueen [12]; Bezdek [14]) search for partition of X into k clusters that minimizes the objective function J with unknown varibles U and C as follows:

$$J(U, C) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} u_{i,l}\ d(x_{i,j}, c_{l,j}) \tag{1}$$

Subject to

$$\sum_{l=1}^{k} u_{i,l} = 1, \text{for } 1 \leq i \leq n, \tag{2}$$

where U is an $n \times k$ partition matrix, $u_{i,l}$ is 0 and 1,$u_{i,l} = 1$ indicates that object i is allocated to cluster l; $C = \{C_1, C_2, ..., C_k\}$ is a set of k vectors representing the centroids of the k clusters; $d(x_{i,j}, c_{l,j})$ is distance between object i and the centroid of cluster l on the jth feature. The distance is euclidean. If the feature is numerical, then

$$d^2(x_{i,j}, c_{l,j}) = \sum_{j=1}^{m} (x_{i,j} - c_{l,j})^2 \tag{3}$$

If the feature is categorical, then $X_i = (x_{i,1}, x_{i,2}, ..., x_{i,m})$ for

$$x_{i,j} = w_{i,j} y_{i,j}, \text{for } j = 1, 2, ..., m \tag{4}$$

Subject to

$$\sum_{j=1}^{m} w_j = 1, \text{for } 1 \leq w_j \leq i, \tag{5}$$

where y is categorical feature, and w is weighting.

The above optimization problem can be solved by iteratifely solving the following two minimization problems:

1. Fix $C = \hat{C}$ and solve the reduced problem $J(U, C)$. Problem J_1 is solved by

$$\begin{cases} u_{i,l} = 1, \text{if } \sum_{j=1}^{m} d(x_{i,j}, c_{l,j}) \leq d(x_{i,j}, c_{t,j}), \text{for } 1 \leq t \leq k \\ u_{i,t} = 0, t \neq l \end{cases} \tag{6}$$

2. Fix $U = \hat{U}$ and solve the reduced problem $J(U, C)$. Problem J_2 is solved by

$$c_{l,j} = \frac{\sum_{i=1}^{n} u_{i,l} x_{i,l}}{\sum_{i=1}^{n} u_{i,l}}, \text{for } 1 \leq l \leq k, \text{and } 1 \leq j \leq m. \tag{7}$$

Step 2. After obtaining clusters from step 1, and then to make new clusters based on the rules. Let $K = \{K_1, K_2, ..., K_k\}$ be a set of k clusters from step 1 where cluster $K_i = \{X_1, X_2, ..., X_i\}$, the new cluster $G_i = \{X_1, X_2, ..., X_i\}$ is obtained by rules in Table 1. In the result of Step 2, we have end clusters.

Table 1. Rules

	Clusters K		New cluster G
If	$K_1, K_2, ..., K_k$	**Then**	G_i

3. RESULT AND DISCUSSION

The data of simulation contains information on items to be sent with a specific destination address. Data consists of 102 records and presented in Table 2. After each item is clustered based on features, then items are grouped for delivery using the type of item delivery vehicle based on rules.

This data set contains 2 numerical and 1 categorical. The features are standardized by Z_{score}. The results of processing data with rule-k-means are; features are clustered into 3 clusters; initial centroid of clusters are determined randomly, that are shown in Table 3; the clusters results for each variable are shown in Table 4 and Table 5; the centroid of clusters are shown in Table 6.

Table 2. Summary of simulation data

Features	Type of Features	Label
Weight	Numerical	-
Volume	Numerical	-
Type of road	Categorical	Small way (gang)
		Highway traffic
		Highway
		Main highway

Table 3. Intial centroid

Clusters	Weight	Volume	Type of road
K_1	3.36	2.80	-0.63
K_2	0.64	1.78	3.17
K_3	-0.51	-0.53	1.37

Table 4. Clusters in weight and volume

K_1	K_2	K_3
X_1	$X_{11}X_{59}$	$X_2X_{15}X_{23}X_{33}X_{42}X_{54}X_{64}X_{74}X_{81}X_{91}X_{99}$
X_5	$X_{39}X_{61}$	$X_3X_{16}X_{24}X_{34}X_{45}X_{56}X_{65}X_{75}X_{82}X_{92}X_{100}$
X_8	$X_{40}X_{67}$	$X_4X_{17}X_{25}X_{35}X_{47}X_{57}X_{66}X_{76}X_{83}X_{93}X_{101}$
X_{10}	$X_{43}X_{70}$	$X_6X_{18}X_{26}X_{36}X_{48}X_{58}X_{68}X_{77}X_{84}\ X_{95}X_{102}$
X_{12}	$X_{44}X_{87}$	$X_7X_{19}X_{27}X_{37}X_{49}X_{60}X_{69}X_{78}X_{85}X_{96}$
X_{21}	$X_{46}X_{90}$	$X_9X_{20}X_{30}X_{38}X_{50}X_{62}X_{71}X_{79}X_{86}\ X_{97}$
X_{28}	$X_{52}X_{94}$	$X_{13}X_{22}X_{31}X_{41}X_{51}X_{63}X_{72}X_{80}X_{88}X_{98}$
X_{29}	X_{55}	$X_{14}X_{32}X_{53}X_{73}X_{89}$

The clusters in weight and volume, cluster 1 shows the characteristics of the most heavy items, while cluster 2 shows characteristics of items with a weight and volume smaller than cluster 1, and cluster 3 shows characteristics of items that weight and volume smaller than cluster 2. From 102 data, cluster 1 consists of 8 objects, cluster 2 consists of 14 objects, and cluster 3 consists of 80 objects.

Table 5. Clusters in type of road

K_1	K_2	K_3
$X_1X_{12}X_{20}X_{34}X_{47}X_{57}X_{64}X_{76}X_{84}$	X_{10}	$X_5X_{39}X_{74}X_{96}$
$X_2X_{13}X_{21}X_{40}X_{48}X_{58}X_{65}X_{77}X_{85}$	X_{11}	$X_{16}X_{41}X_{82}X_{97}$
$X_3X_{14}X_{26}X_{42}X_{49}X_{59}X_{66}X_{78}X_{85}$	X_{22}	$X_{23}X_{53}X_{87}X_{99}$
$X_4X_{15}X_{27}X_{43}X_{50}X_{60}X_{67}X_{79}X_{91}$	X_{24}	$X_{25}X_{55}X_{88}X_{100}$
$X_6X_{17}X_{28}X_{44}X_{51}X_{61}X_{69}\ X_{80}X_{92}$	X_{29}	$X_{35}X_{68}X_{90}\ X_{102}$
$X_7X_{18}X_{31}X_{45}X_{52}X_{62}X_{71}X_{81}X_{93}$	X_{30}	$X_{36}X_{70}X_{98}$
$X_8X_{19}X_{32}\ X_{46}X_{54}X_{63}X_{72}X_{83}X_{94}$	X_{86}	$X_{37}X_{73}X_{101}$
$X_9X_{33}X_{56}X_{75}X_{95}$		X_{38}

The clusters in type of road, cluster 2 shows the type of road with the largest weight, while cluster 3 shows the type of road with a weight smaller than cluster 2, and cluster 1 shows the type of road which is smaller than cluster 2. From 102 data, cluster 1 consists of 73 objects, cluster 2 consists of 7 objects, and cluster 3 consists of 22 objects.

Table 6. Centroid

Clusters	Weight	Volume	Type of road
K_1	3.26	2.25	-0.53
K_2	0.13	1.55	3.06
K_3	-0.36	-0.52	0.79
Iteration	2	2	3

In this case, iteration of data clustering occurs in 2 iterations of cluster in weight and volume, 2 times iterations of clusters in type of road. In these iterations, the centroid of each cluster has not changed and there is no more data moving from one cluster to another.

For labeling the type of vehicle shown in Table 7. From the results of the clustering, a decision was made to determine the type of item delivery vehicles with the rules shown in Table 8.

Table 7. Type of item delivery vehicles

Vehicles	Cluster
Large box car	G_1
Small box car	G_2
Motorcycle	G_3

In this case, we have 9 rules for mapping item delivery. Example: if we have items with first cluster in weight and volume, and second cluster in type of road, then the item delivery use a *large box car.* All rules are shown in Table 8.

Table 8. Type of item delivery vehicles

Cluster in weight and volume	Cluster in type of road	New cluster G_i
1	2	G_1
1	3	G_1
1	1	G_3
2	2	G_1
2	3	G_2
2	1	G_3
3	2	G_1
3	3	G_2
3	1	G_3

Based on the regulation, a decision is obtained in Table 9.

Table 9. Decision of Item delivery vehicles

Large box car G_1	Small box car G_2	Motorcycle G_3
X_5	$X_{16}X_{41}X_{82}$	$X_1X_{12}X_{20}X_{34}X_{47}X_{57}X_{64}X_{76}X_{84}X_{96}$
X_{10}	$X_{23}X_{53}X_{87}$	$X_2X_{13}X_{21}X_{40}X_{48}X_{58}X_{64}X_{77}X_{85}X_{97}$
X_{11}	$X_{25}X_{55}X_{88}$	$X_3X_{14}X_{26}X_{42}X_{49}X_{59}X_{66}X_{78}X_{89}X_{99}$
X_{22}	$X_{35}X_{68}X_{90}$	$X_4X_{15}X_{27}X_{43}X_{50}X_{60}X_{67}X_{79}X_{91}X_{100}$
X_{24}	$X_{36}X_{70}X_{98}$	$X_6X_{17}X_{28}X_{44}X_{51}X_{61}X_{69}X_{80}X_{92}X_{102}$
X_{29}	$X_{37}X_{73}X_{101}$	$X_7X_{18}X_{31}X_{45}X_{52}X_{62}X_{71}X_{93}$
X_{30}	$X_{38}X_{74}$	$X_8X_{19}X_{32}X_{46}X_{54}X_{63}X_{72}X_{81}X_{94}$
X_{86}	X_{39}	$X_9X_{33}X_{56}X_{75}X_{83}X_{95}$

The final decision obtained information that the large box car will deliver 8 objects, 21 objects will be delivered by the small box car, and 73 objects will be delivered by motorcycle.

4. SUMMARY AND FUTURE WORK

Quantification of categorical data through weighting results is better cluster characteristics than using the rank approach, and converts it into binary. The combination of k-means and rules (rule-k-means) to obtain a new cluster makes it easier to group objects according to the desired characteristics or objectives. For the grouping, the practice of item delivery services companies using clustering methods and then use rule-k-means algorithm to determine the types of vehicles to be used. In future works, we will work on rule-k-means algorithm with traveling salesman problem for effective route.

Acknowledgement. This research is mainly supported by BOPTN ITB.

REFERENCES

[1] Trivedi, S., Pardos, Z. A., Sarkozy, G. N. and Heffernan, N. T., "Spectral clustering in educational data mining", *Proceedings of International Conference on Educational Data Mining*, **4** (2011), 138-147.

[2] Ralambondrainy, H., "A conceptual version of the k-means algorithm", *Pattern Recognition Letters*, **16** (1995), 1147-1157.

[3] Huang, Z., "Extensions to the k-means algorithm for clustering large data sets with categorical values", *Data Mining and Knowledge Discovery*, **2** (1998), 283-304.

[4] Huang, J. Z., Ng, M. K., Rong, H. and Li, Z., "Automated variable weighting in k-means type clustering", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **27** (2011), 657-667.

[5] Celebi, M. E., "Improving the performanc of k-means for color quantization", *Image and Vision Computing*, **29** (1998), 260-271.

[6] Dhanachandra, N., Manglem, K. and Chanu, Y. J., "Image segmentation using k-means algorithm and subtractive clustering algorithm", *Eleventh International Multi-Conference on Information Processing*, **54** (2015), 764-771.

[7] Sahu, B., Dehuri, S. and Jagadev, A. K., "Feature selection model based on clustering and ranking in pipeline for microarray data", *Informatics in Medicine Unlocked*, **9** (2017), 107-122.

[8] Everitt, B. S., Landau, S., Leese, M., and Stahl, D., *Cluster analysis*, 5^{th} edition, John Wiley and Sons, 2011.

[9] Oliveira, J. V. and Pedrycz, W., *Advances in fuzzy clustering and its aplication*, John Wiley and Sons, 2007.

[10] Sugiyama, M., *Introduction to statistical learning machine*, Morgan Kaufman and Elsevier, 2016.

[11] Anderberg, M. R., *Cluster analysis for applications*, Academic Press, 1973.

[12] MacQueen, J. B., "Some methods classification and analysis of multivariate observations", *Proceedings of the 5^{th} Berkeley Symposium on Mathematical Statistics and Probability*, **1** (1967), 281-297.

[13] Rencher, A. C. and Christensen., *Methods of multivariate analysis*, 3^{rd} edition, John Wiley and Sons, 2012.

[14] Bezdek, J. C., *Pattern recognition with fuzzy objective function*, Plenum Press, 1981.