Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 72 (2015) 495 – 502

Procedia Computer Science

The Third Information Systems International Conference

Grouping of Retail Items by Using K-Means Clustering

Kusrini Kusrini*

STMIK AMIKOM Yogyakarta, Jl.Ringroad Utara Condong Catur, Depok Sleman Yogyakarta, 55283, Indonesia

Abstract

Two main activities in retail business are to determine the amount of stock that should be maintained and the profit margin for each item. As the inventory principle stated, both processes required to make categories group of 'fast moving' and 'slow moving' and use it as the consideration for the processes. Citramart Minimarket as a retail business has not yet been used that grouping and consideration in their sales information system for processing their item's minimum stock level and profit margin. The process is still done manually by arbitrary observations from a stock division staff. The categories that is used to determine the minimum stock and profit margin are also not the moving speed of item but rather the kind of items.

This study aim to support the process of determining the minimum stock and profit margin by building a model that can group items into categories 'fast moving' and slow moving' using k-means clustering. K-means clustering is used in this study because the number of clusters required in categorization of items already set. The group cluster which has highest centroid will be the fast moving group, while the lowest centroid is the slow moving group. The data to be used in the research is taken from sales data for year 2014 and 2015. The clustering scenario uses combination of time, e.g. yearly and monthly, and variable, e.g. count of items and their transaction values. Using 3 clusters and delta value 0.2, the resulting best scenario is using yearly time and transaction value variable. The test is conducted by calculating the xie-beny index which results 36.265.

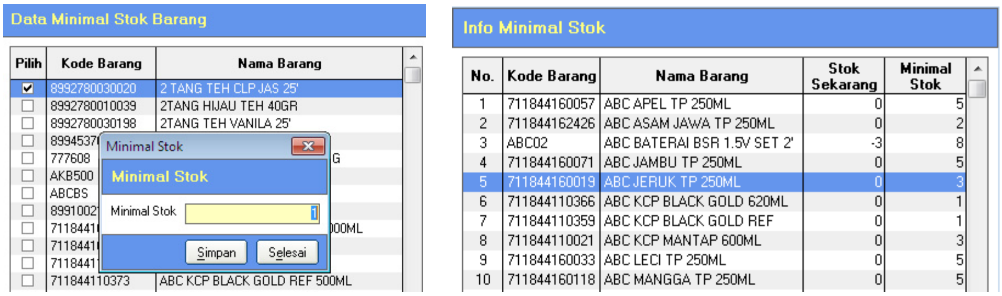
Keywords: k-means clustering, retail, minimal stock, profit margin;

1. Introduction

Citramart is a minimarket in STMIK AMIKOM Yogyakarta. Citramart serves the need of more than 10.000 students and 200 employees. The number of items to sell in Citramart is 5.492. Those items is grouped into several categories such as stationary, drink, snack, household, accessories, etc. Each item minimum stock level is defined in the item management page of sales software system. This data then be used to warn the minimarket staff about items which stock's is falls below the minimum level. The

doi:10.1016/j.procs.2015.12.131

software's user interface for setting the minimum stock and warning the under stock limit items are shown in Figure 1.



Data Minimal Stok Barang

Pilih	Kode Barang	Nama Barang
☑	8992780030020	2 TANG TEH CLP JAS 25'
□	8992780010039	2TANG HIJAU TEH 40GR
□	8992780030198	2TANG TEH VANILA 25'
□	8994537	
□	777608	G
□	AKB500	
□	ABCBS	
□	8991002	
□	7118441	00ML
□	7118441	
□	7118441	
□	711844110373	ABC KCP BLACK GOLD REF 500ML

Minimal Stok

Minimal Stok

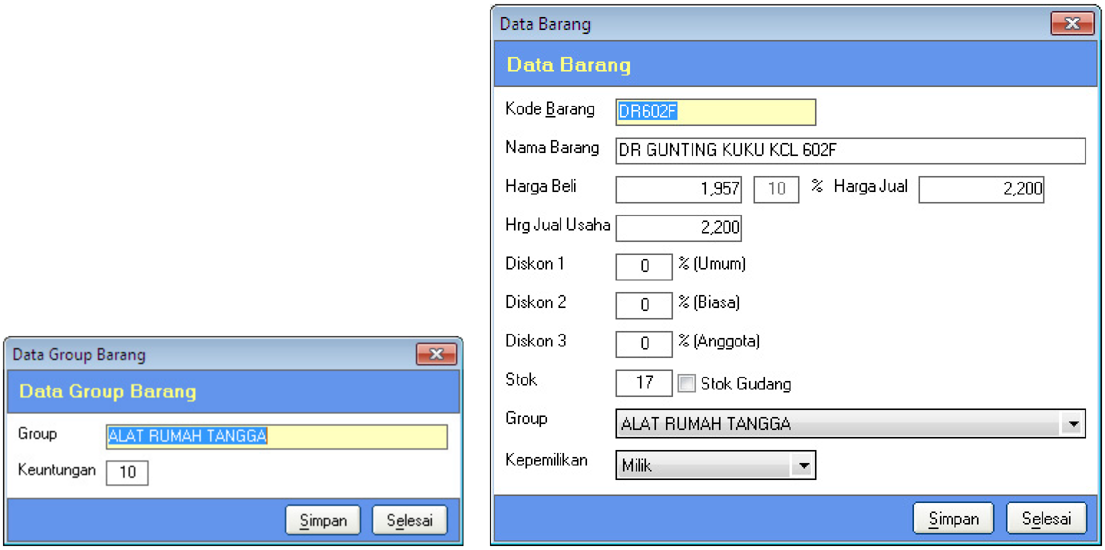
Minimal Stok: 1

Info Minimal Stok

No.	Kode Barang	Nama Barang	Stok Sekarang	Minimal Stok
1	711844160057	ABC APEL TP 250ML	0	5
2	711844162426	ABC ASAM JAWA TP 250ML	0	2
3	ABC02	ABC BATERAI BSR 1.5V SET 2'	-3	8
4	711844160071	ABC JAMBU TP 250ML	0	5
5	711844160019	ABC JERUK TP 250ML	0	3
6	711844110366	ABC KCP BLACK GOLD 620ML	0	1
7	711844110359	ABC KCP BLACK GOLD REF	0	1
8	711844110021	ABC KCP MANTAP 600ML	0	3
9	711844160033	ABC LECI TP 250ML	0	5
10	711844160118	ABC MANGGA TP 250ML	0	5

Fig.1. (a) Interface to determine minimum stock; (b) Information of items below minimum stock

The system set an item's profit margin based on its category. In other words, the Citramart staff has to set the category to be used for the item first and the profit margin for the category. The category and profit margin data that has been input will then be used as basic to set the selling price when input the item's purchasing data. The user interfaces to set a category's profit margin and to inform selling price from its category's profit margin are shown in Figure 2.



Data Group Barang

Data Group Barang

Group: ALAT RUMAH TANGGA

Keuntungan: 10

Data Barang

Data Barang

Kode Barang: DR602F

Nama Barang: DR GUNTING KUKU KCL 602F

Harga Beli: 1,957 | 10 % | Harga Jual: 2,200

Hrg Jual Usaha: 2,200

Diskon 1: 0 % (Umum)

Diskon 2: 0 % (Biasa)

Diskon 3: 0 % (Anggota)

Stok: 17 □ Stok Gudang

Group: ALAT RUMAH TANGGA

Kepemilikan: Milik

Fig.2. (a) Setting profit margin for category (Group); (b) Information of profit margin and selling price based on category (Group)

By using the method above, defining items' minimum stock and its profit margin is only based on staff estimation. This is not conform with inventory principles that defining process of minimum stock must consider whether an item is in category of fast moving or slow moving.

Fast moving is an item category in the warehouse that has a fast transaction number or used frequently in every period. In opposite way, slow moving is an item category in warehouse with rarely transaction. A stop moving is a category for zero out transaction or unused item.

Researches about determining safety stock has been conducted by Ocrun[1] and Kucer[2]. A research that is discussed about determining item category in fast moving has conducted by Srivastav [3] when Giering[4] has conducted a research that utilized clustering in retail area. In this research, determining item's group into categories of fast moving and slow moving in retail industry is done by using clustering process. Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters[5].

The K-means is one of algorithm that commonly used in clustering process is K-means clustering. The "K" in its name refers to the fact that the algorithm looks for a fixed number of clusters which are defined in terms of proximity of data points to each other [6].

This research's result is expected to be a reference for Citramart Minimarket for improving their method and information system to determine the minimum stock and profit margin by following the new improved grouping model. Moreover, this items grouping model is also expected to be used as a reference for other retails in general, as support for their various decision making needs.

2. Research Method

2.1. Research Flow

This research will conduct an item cluster formation simulation based on sales data in order to determine which items that is going into fast moving or slow moving cluster. The data will be clustered to 3 clusters, the highest centroid value cluster will be labeled as fast moving item group, while the lowest centroid value will be labeled as slow moving item group. The clustering process only applied to item data that have transaction count > 0, or in other words, not a stop moving item category.

The clustering process will be conducted in several scenarios. The compactness value and distance between cluster for each scenarios will be calculated using xie-beny index. Clustering process with the highest xie-beny index will be recommended to be used in Citramart Minimarket.

The research flow is shown in Figure 3.

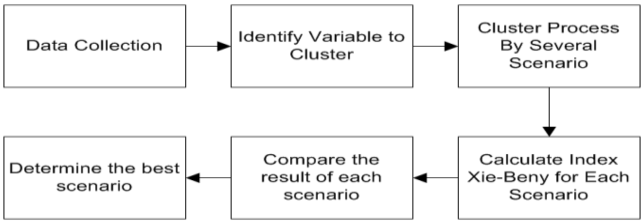

Data Collection
Identify Variable to Cluster
Cluster Process By Several Scenario
Calculate Index Xie-Beny for Each Scenario
Compare the result of each scenario
Determine the best scenario

Fig.3. Research Flow

2.2. Data and Variable

The data that is used as the clustering object is the sales data of Citramart Minimarket for year 2013 and 2014. The data are grouped into yearly and monthly data. For yearly data, all the data from 2013 and 2014's data will be used. For the monthly data, the data is taken from January 2014, May 2014 and September 2014. The variable to be used is transaction item count and sales value count.

2.3. Cluster Process

Cluster process is using k-means cluster. The input of this process is initial data and delta. Initial data is the sales data according to the variables used for each scenario. Delta is the value to be used to determined the allowed gap between the centroid and mean. In this study, the delta is set at 0.2.

K-Means method is commonly done with basic algorithm as follow [7]:

1. Determine count of class (cluster)
2. Determine initial centroid of each class
3. Put each data into class which has the nearest centroid
4. Calculate the data mean from each class
5. For all class, if the difference of the mean value and centroid goes beyond tolerable error, replace the centroid value with the class mean then go to step 3.

In fundamental k-means clustering algorithm, initial centroid value from specific discrete class is defined randomly, while in this research the value is produced from an equation shown in Equation 1.

$$c_i = \min + \frac{(i-1)*(\max - \min)}{n} + \frac{(\max - \min)}{2*n} \tag{1}$$

Where

c_i : centroid class i

min : the lowest value of continue class data

max : the biggest value of discrete class data

n : total number of discrete class

The building process of discrete class is shown as follow:

1. Specify the source data
2. Specify desired total number of cluster (n)
3. Get the lowest value from source data (min)
4. Get the highest value from source data (max)
5. Specify delta (d) to get the acceptable error by Equation 2

$$e = d*(\max - \min) \tag{2}$$

6. For each cluster, find the initial centroid (c) by Equation 1.
7. For each data, find distance (s_{ij}) between data (d_i) and the centroid of each cluster (c_j)by Equation 3

$$s_{ij} = | d_i - c_j | \tag{3}$$

8. For each value in source data, put it into its appropriate cluster, which is one that has nearest centroid to the value.
9. Calculate the average value of all members for each class`(avg)
10. For each discrete class, calculate the difference between its mean and its centroid (s) by Equation 4.

$$s = \sum_{i=1}^{n} | avg_i - c_i | \tag{4}$$

11. For each discrete class, if s>e , then replace its centroid value with its mean, put out all values from their corresponding class then go back to step 7.
12. Finish.

For more complete example, it is given 10 sales data in order : 10, 9, 15, 100, 225, 75, 60, 9, 8, 76. Those data value will be distributed into 3 cluster with delta 0,01. From the data, it can be evaluated that the min and max values are 8 and 225 respectively.

The acceptable error (e) is produced by applying calculation as follow :

$$e = 0.01 + (225 - 8)$$

$$e = 2.17$$

The first discrete class' centroid (c1) is produced as follow:

$$c_1 = 8 + \frac{(1-1)*(225-8)}{3} + \frac{(225-8)}{2*3}$$

$$c_1 = 44{,}17$$

With the same way, the value c2 and c3 is produced as 116,5 dan 188,83. The calculation of distance between a data value to class' centroid which makes the decision about which class that the value must go into (step 7). Each data value then put into a class which has nearest centroid to the value. The result is shown in Table 1.

Table 1. Cluster Result for 1st Iteration

Data	Distance to Centroid			Cluster
	1	2	3	
10	34.17	106.50	178.83	1
9	35.17	107.50	179.83	1
15	29.17	101.50	173.83	1
100	55.83	16.50	88.83	2
225	180.83	108.50	36.17	3
75	30.83	41.50	113.83	1
60	15.83	56.50	128.83	1
9	35.17	107.50	179.83	1
8	36.17	108.50	180.83	1
76	31.83	40.50	112.83	1

From Table 1, the next step is to find the average value for each cluster and the difference between the average and its centroid. The result is shown in Table 2.

Table 2. Calculation of Average for 1st Iteration

Cluster	Average	Centroid	|Average-Centroid|
1	44.17	31.04	13.13
2	116.50	55.83	60.67
3	188.83	180.83	8.00
		Total	**81.79**

Table 2 also shows that the sum of difference between average and centroid for each cluster is more than acceptable error value. Thus, the centroid value of all clusters are replaced with their average value and the process has to be continued to next iteration of putting in the value to nearest centroid's cluster and checking the error value.

2.4. Xie-Beny Index Calculation

The classes resulted from the clustering process need to be validated. The indicator of clustering evaluation result can be in a form of compactness level and separation level. A small compactness and separation level indicate a good cluster [8]. In this research, validation is done by using xie-beny index, that is a compactness and separation validation function of fuzzy clustering. The xie-beny index value is produced from Equation 5[8].

$$XB = \frac{\sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^{2} \left\| V_i - X_j \right\|^2}{n \underbrace{\min}_{i \neq k} \left\| V_i - V_k \right\|^2} \tag{5}$$

where
XB : xie-beny index
c : number of class
n : number of data
Vi : ith class centroid
Xj : jth Data

The xie-beny index value from Equation 5 is used for fuzzy clustering. In this research, k-means clustering method that have membership value 0 and 1 is used. Therefore, the xie-beny index formula is adapted as shown in Equation (6).

$$XB = \frac{\sum_{i=1}^{c} \sum_{j=1}^{mc} \left\| V_i - X_j \right\|^2}{n \underbrace{\min}_{i \neq k} \left\| V_i - V_k \right\|^2} \tag{6}$$

where
XB : xie-beny index
c : number of class
n : number of data
mc : number of data in class c
Vi : ith class centroid
Xj : jth Data

For collection with less than 10 data, the determination of number of class is done by doing XB calculation process from the 2nd data to (n-1)th data, while for collection with more than 10 data, the calculation process is done from the 2nd data to (n/3)th data. The process will be terminated instantly whenever the condition XB(c) ≤ XB(c-1) is met.

3. Result and Discussion

3.1. Result

The cluster centroid and xie-beny index of 10 clustering process is shown in Table 3, while the average of xie-beny index for each scenario is shown in Table 4.

Table 3. Centroid and Index Xie-Beny for each Cluster Process

No	Data	Variable	C.C1	C.C2	C.C3	Xie-Beny Index
1	2013	V1	2191	167632	363688	3.4
2	2014	V1	2386	221198	652517	2.27
3	Jan-14	V1	324	0	44498	0.99
4	May-14	V1	538	31711	85709	2.49
5	Sep-14	V1	458	23605	50977	3.4
6	2013	V2	19607683	58804050	98000416	67
7	2014	V2	3779098	55597145	93582427	5.53
8	Jan-14	V2	1154946	36374075	84690600	2.98
9	May-14	V2	1560066	38321914	74758033	4.02
10	Sep-14	V2	1432749	34407280	64492333	4.37

Note:

V1: Number of Items, V2: Transaction Value

C.C1: Centroid of Cluster 1, C.C2: Centroid of Cluster 2, C.C3 Centroid of Cluster 3

Table 4. Centroid and Index Xie-Beny for each Cluster Process

No	Data	Variable	Xie-Beny Index Average
1	Yearly	V1	2.835
2	Monthly	V1	2.293333
3	Yearly	V2	36.265
4	Monthly	V3	11.37

Based on Table 4, it is shown that the best cluster with xie-beny index 36.265 is occurred in scenario data yearly with transaction value variable.

4. Conclussion

K-Means clustering can be used in item grouping process into categories of fast moving and slow moving. By using the sales data in Citramart Minimarket of STMIK AMIKOM Yogyakarta year 2013 and 2014, it is shown that the best cluster is occurred in clustering process with yearly data and variable transaction value. The value of xie-beny index for that cluster is 36.265

5. Further Work

In the future research, it is possible to add testing scenario with the number of transaction variable. Beside that, it can be tried to do clustering by using more than 1 variable.

References

[1] Orcun, S., Çetinkaya, S and Uzsoy, R., Determining safety stocks in the presence of workload-dependent lead times. *Proceedings of the 2007 Winter Simulation Conference*, 2007;p.1691-1698

[2] Kocer, U.U., Tamer, S., Determining the Inventory Policy for Slow-Moving Items: A Case Study. P*roceedings of the World Congress on Engineering 2011 Vol I, WCE 2011, July 6 - 8, 2011, London, U.K;p.139-143.*

[3] Srivastav, A., Agraval, s., Multi Objective Cuckoo Search Optimization for Fast Moving Inventory Items, Advances in Intelligent Informatics, *Springer*, 2015, DOI 10.1007/978-3-319-11218-3

[4] Giering, M, Retail sales prediction and item recommendations using customer demographics at store level, *SIGKDD Explorations*, Volume 10 Issue 2, ACM New York, NY, USA, 2008, 10.1145/1540276.1540301. p84-89

[5] Phrabu, S., Venatesan, N. Data Mining And Warehousing, *New Age International Publisher*, 2007, p.34-40

[6] Berry, M.J.A. dan Linoff, G. S.. Data Mining Techniques For Marketing, Sales, and Customer RelationshipManagement, Second Edition*, Wiley Publishing, Inc*., Indianapolis, Indiana. 2004,

[7] Larose, D. T., Discovering Knowledge in Data: An Introduction to Data Mining. *John Wiley and Sons*, 2005. pp: 116-126 and 153-158. ISBN: 0471666572. DOI: 10.1002/0471687545

[8] Xie, X.L. dan Beni, G., A Validity Measure for Fuzzy Clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1991, 8, 13, 841-847.