ANALYSIS OF HIERARCHICAL CLUSTERING AND K-MEAN METHODS WITH LRFMP MODEL ON CUSTOMER SEGMENTATION

ASEP MUHIDIN

ABSTRACT

Customer is something valuable and important, if all customers are similar, business will be so simple. The problem of heterogeneity and the large number of customers is a challenge to be faced in determining potential customer segmentation. Companies must be able to plan, create and implement strategies for treating heterogeneous consumer traits. RFM Model Approach, which is a segmentation model based on the attributes of Recency, Frequency, and Monetary. Model RFM is a segmentation model commonly used in companies.

In this research, customer segmentation process begins with preprocessing process, analytic hierarchy process (AHP), search for the best value of all Hierarchical Clustering methods by comparing the Bouldien-Index value. Furthermore, the value of K is chosen to be the initial value in K-Mean Clustering. The clustering result is used to segment the RFM model to get the consumer class. The addition of Payment parameter (LRFMP) can increase the value of customer loyalty to the company.

Based on the research results, the single linkage method is the best method to find the value of K. Segmentation of the k-mean model with the addition of the P (LRFMP) parameter can increase the DBI value compared to the weighted RFM model or not. But the DBI value of the single linkage segmentation method is still better than the k-mean segmentation.

Keywords : CRM, data mining, preprocessing, Hierarchical Clustering, Bouldien-Index, clustering, segmentation, RFM, LRFMP, Customer.

1. Introduction

Many organizations have come to the conclusion that understanding of customers is valuable and important. If all customers are similar, the business will be so simple. A good segmentation analysis of customers can be used to map potential and potential customers.

Customer transaction data is growing very rapidly, so the data volume is getting bigger and bigger day, either from the number of record and number of field. With the development of such data it is not possible to analyze customer behavior with traditional or manual methods.

The use of data mining methods can be used to analyze data obtained from transactions, so as to extract hidden information from the transaction data, and one of them is customer segmentation information. There are two clustering methods, namely Hierarchical Clustering and Non Hierarchical Clustering. The Hierarchical Clustering Method consists of Single Linkage Clustering, Complete Linkage Clustering, Average Linkage Clustering, Centroid Linkage Clustering and Ward methods. While the method of Non Hierarchical Clustering itself consists of K-means and Fuzzy K-means.

In this research will be used all methods of Hierarchical Clustering to determine the best number of clusters. Furthermore the best number of clusters is used for the value of K in customer clustering process with K-Mean method. Customer segmentation uses RFM parameters, the customers are grouped by value of Recency, Frequency and Monetery. Recency is the last time the customer makes a transaction with the time of research in units of months. Frequency is the number of transactions within the study period. While Monetery is the total value of transactions within the period of study. Segmentation results can determine customer ratings and customer profiles.

Based on previous research data, no customer segmentation has been done using Hierarchical Clustering and K-Means method with LRFMP model, so this research will analyze Hierarchical Clustering and K-Means method on the characterization of customer based on LRFMP model (Length, Recency, Frequency, Monetary ,

Payment). Model P (Payment) researcher is different from that already done by A. Parvaneh, M. J. Tarokh and H. Abbasimehr namely P (Potential).

2. LiteratureReview

2.1 Hierarchical Methods

Hierarchical methods are one of the clustering techniques by grouping two or more objects that have the closest resemblance value. Furthermore the first grouping results are grouped again with another object that has a value of similarity second. And so on, so it will form a hierarchical construction or based on certain levels such as tree structure (game structure). Dendogram is used to describe the structure of the cluster hierarchy results.Some Methods The hierarchical cluster technique used is as follows:

1) *Agglomerative Methods*

The Agglomerative Methods method approach begins by defining each data object to form its own cluster. The next process is to combine two objects that have the nearest distance value and form a new cluster. Subsequently another object will join the new cluster or along with other objects forming a new cluster.

Some techniques in Agglomerative methods are:

a) Single Linkage (Nearest Neighbor Methods)
b) Complete Linkage (Furthest Neighbor Methods)
c) Average Linkage Methods (Between Groups Methods)

2) Centroid Method

In the method of centroid method, the distance between two clusters formed is the distance between the two centroids in the two clusters. Centroid is the average distance between objects in a cluster. The value of the centroid is obtained by doing an average on all cluster members.

3) Ward's Method

Ward method does not calculate the distance between clusters. However, forming clusters by maximizing homogeneity in a cluster. Homogeneity is measured by using the sum of the squared deviations in the mean cluster for each observation. Error sum

of squares (SSE) is used as an objective function. Two objects will be combined if they have the smallest objective function among possibilities.

$$SSE = \sum_{j=1}^{p} (\sum_{i=1}^{n} X_{ij}^2 - \frac{1}{n} (\sum_{i=1}^{n} X_{ij})^2)$$

Where :

Xij : value for the i th object of the jth cluster

P : number of variables measured

N : the number of objects in the cluster formed

4) K-Mean Cluster

K-Means method is one non-hierarchical clustering method that performs unsupervised modeling process and is one of the methods that do the data grouping with partition system. This method performs group analysis that leads to the division of the observed object into the cluster, this method tries to find the center of the group (centroid) in the data as much as the iteration of the improvement performed. This method attempts to divide the data into clusters so that the same characteristic data is entered into one group, while different characteristic data are included in the other group.

2.2 Indeks Davies-Bouldin (IDB)

The Davies-Bouldin (IDB) index is a cluster validation method for the quantitative evaluation of clustering results. This method aims to maximize the measurement of cluster distance between one cluster with another cluster, also minimize the distance between members in one cluster. In this study IDB will be used to detect outliers on each cluster formed.

$$Var\,(x) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

Where :

$\bar{x}$: The average of cluster x and

N: number of cluster members.

Then calculate the Davies-Bouldin Index (DBI) with the equation

$$DBI = \frac{1}{k}\sum_{i=1}^{k} R_i$$

$$R_i = \max_{j=1..k, i \neq j} R_{ij}$$

$$R_{\substack{ij \\ i \neq j}} = \frac{\text{var}(C_i) + \text{var}(C_{j)}}{\|c_i - c_j\|}$$

Where :

Ci = cluster i and ci are the centroid of cluster I

2.3 Analytic Hierarchy Process (AHP)

Prof. Thomas L.Saaty developed a decision-making algorithm method to solve multicriteria problems. This method is known as Analytic Hierarchy Process or AHP method. AHP simplifies the multicriteria problem into a hierarchy of three main components: the goal or goal of decision making, assessment criteria and alternatives. The priority set of elements table with pairwise comparisons is used for comparative scales as measures that state the intensity of interest (Saaty 1988).

Intensity of Interests	Description	Explanation
1	Both elements are equally important	Two elements have the same effect on the goal
3	One element is slightly more important than the other elements	Experience and judgment support a little more than any other element
5	One element is more important than the other	Experience and judgment are very strong in favor of one

		element over the other
7	One element is clearly more important than the other elements	One powerful element is emptied of the dominant sanity seen in practice
9	One element is absolutely essential from other elements	Evidence that supports one element against other elements has the highest degree of affirmation that might be corroborating
2,4,6,8	Values between two adjacent consideration values	This value is given when there are two compromises between two options
The opposite	If for activity i gets one number compared with activity j, then j has the opposite value compared with i.	

3. Research Methodology

In this research methodology section will be described systematic and directed steps that will be used as a framework of customer segmentation research using all methods of Hierarchical Clustering and K-Mean with LRFMP (Length Recency, Frequency, Monetary, Payment) variables, Which produces the best cluster results.

3.1 Information Data

At this stage an observation and understanding of available data for the data mining process, comprising collecting and verifying data quality. The company database includes a wide range of data. Among the existing data, there are customer data and customer transactions that meet the research needs for customer segmentation.

3.2 Pre-processing Data

Preprocessing data is an important step in the data mining process, as it improves the accuracy and efficiency of subsequent modeling. In this research the preprocessing data is done:

- Data Selection
- Data *Preprocessin*
- Data *Transformation*
- Standarisasi Data

3.3 RFM Modeling

RFM model is a model commonly used to assess customer loyalty, this model was introduced by Highes (1994), which is built on three criteria:

a) Recency: refers to the duration of the current / last purchase period with the current; A lower value corresponds to a higher probability of a customer making a repeat purchase.
b) Frequency: the number of purchases made within a certain time / period; Higher frequencies show greater loyalty.
c) Monetary: the total value of purchases spent during a given period; Higher numbers indicate a large contribution to the company.

In this study the authors propose the addition of a new variable that is P (Payment). P is the amount of overdue (overdue payments) made by the customer. Timely payment is a mis-measure of customer loyalty to the company.

3.4 Analytical Hierarchy Process (AHP) for Weighting of LRFMP Model

The AHP method is used to determine the priority weight of the LRFMP model criteria denoted by wL, wR, wF, wM and wP. The AHP method is implemented in the marketing party's evaluation results in assessing the influence of criteria of length, recency, frequency, monetary, and payment on company operations. The next step calculates the inconsistency index value and is checked for each assessment, and finally we get the weight value of each criterion. The weight is multiplied by the LRFMP model criterion value as cluster rank determinant with Equation:

$$\text{LRFMP weighted} = wl \times L + wr \times R + wf \times F + wm \times M + wp \times P$$

The higher the rank, the greater the level of customer keloyalan in the cluster.

3.5 Determine the optimal K value based on Davies-Bouldin Index

AHP result data in segmentation with all methods of Hierarchical Clustering, ie from Complete Linkage, Single Linkage, Average Linkage Clustering, Centroid and Wards method. Segmentation result formed will be evaluated using Davies-Bouldin (DB) Index. The smaller the DB Index value indicates the most optimal cluster scheme.

3.6 Clustering with K-Mean Clustering

Clustering stage using K-Mean Clustering algorithm. The data used as input is the normalized LRFM value data with weight and LRFMP normalization with weights. While the initial K value is the K value of the analysis of Davies Bouldin Index.

5. Conclusion

1) Single Hierarchical Clustering Method is the best method for determining K value of initial K-mean method compared with Single Ward method, Complete Linkage, Average Linkage and Centroid
2) Added P parameter (payment) able to make better DBI value.
3) Added weight can increase DBI cluster value.
4) The value of LRFM model accuracy ratio is better than LRFMP model
5) The number of clusters (K) results of the Hierarchical Clustering method, can be applied to the initial K value of k-mean. But with the same number of clusters, the single linkage DBI method is better than the k-mean method.
6) The results of clustering can be used as a marketing reference in determining the treatment of customers.

References

Mohammed J. Zaki and Wagner Meira Jr. 2014.*Data Mining And Analysis : Fundamental Concepts and Algorithms,* Cambridge University Press is part of the University of Cambridge

Pang-Ning Tan, Michael Steinbach dan Vipin Kumar, 2006, *Introduction to Data Mining*. Boston San Francisco New York

Jiawei Han and Micheline Kamber, 2006. *Data Mining: Concepts and Technques, Second Edition*. University of Illinois at Urbana-Champaign.

Subhash Sharma, 1996. *Applied Multivariate Techniques*.

University of South Carolina. New York Chichester Brisbane Toronto Singapore

A.Parvaneh, M. J. Tarokh dan H. Abbasimehr, 2014. *Combining Data Mining and Group Decision Making in Retailer Segmentation Based on LRFMP Variables*. IUST Publication ,IJIEPR, Vol. 25, No. 3, All Rights Reserved

Rachid Ait Daoud, Abdellah Amine, Belaid Bouikhalene, Rachid Lbibb 205, *Customer Segmentation Model in E-commerce Using Clustering Techniques and LRFM Model: The Case of Online Stores in Morocco*. World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:9, No:8, 2015

Duen-Ren Liu, Ya-Yueh Shih,2004 *Integrating AHP and data mining for product recommendation based on custonxer lifetime value*. Elsevier B.V. All rights reserved.

Yohana Nugraheni, 2011. Data Mining Dengan Metode Fuzzy Untuk Customer Relationship Management (CRM) Pada Perusahaan Retail. Tesis Universitas Udayana Denpasar.

Yani Soraya, 2011. *Perbandingan Kinerja Metode Single Linkage, Metode Complete Linkage dan Metotle K-means dalam Analisis Cluster*, Skripsi Universitas Negeri Semarang

Bilson Simamora, 2005. *Analisis Multivariat Pemasaran*. PT Gramedia Pustaka Umum

Thomas L. Saaty, 2008. *Decision Making With The Analytic Hierarchy Process*. University of Pittsburgh,