Analysis of Clustering for Grouping of Productive Industry by K-Medoid Method

Indah Cahya Dewi[1], Bara Yuda Gautama[2], and Putu Arya Mertasana[3]

[1][2]Department of Electrical and Computer Engineering, Post Graduate Program, Udayana University
[3] Department of Electrical and Computer Engineering, Faculty of Engineering, Udayana University
Email: dewaayuindah@student.unud.ac.id

***Abstract*— With the number of existing data, would have difficulty in doing the classification and the classification of the existing data. To resolve the issue, one way to do clustering is with data mining using clustering technique. The purpose of this research is the importance of knowing the pattern of the production of an industry that can provide the decision and the construction of clustering patterns for development and industrial progress. The results of this research can provide recommendations to improve the development of industry, help the owners of industry to develop the industry to an increase in the number of production and product quality, improve the competitiveness of the owner of the industry in developing its products. In this research will use the K-Medoids algorithm for data grouping of the industry so that it will be found the information that can be used for the recommendations of the improvement of marketing. The results of clustering with the number of cluster 3 produces the first group contains 85 members, the second group contains 222 members and the third group numbered 3 members. The third group are classified as productive because it has a combination of the value of the production of the most high the results of clustering have the quality of purity worth 1 means good cluster quality.**

***Index Terms*— Cluster, K-Medoid, Marketing Strategy**

I. INTRODUCTION

Along with the development of technology that increases produce a great amount of data. It trigger to collect and process a large amount of data to a more accurate and faster.

With the development of the data warehouse, data min-ing become very attracted the attention of industry information in the past few years is due to the availability of a large amount of data and the greater the need to change the data become information and knowledge that useful.

Take advantage of the grouping of the data manually or done by the man of course has some limitations especially accommodate the data enough to grouped. Besides that does not close the possibility of an error occurred in the conduct of the grouping of data. To resolve the issue, one way to do clustering is with Data Mining using Clustering technique. In this paper will be discussed about one of the data mining techniques namely clustering.

Data mining is needed in the search for important information from the existing handicrafts data over the years. Through data mining, important patterns of handicraft data obtained. A large amount of data can be analyzed with data mining. Data division of a large amount of handicrafts using clustering technique.

This research uses clustering k-medoid method. This research uses quantitative research method because it uses the data in the Provincial Industry and Trade of Bali Province.

II. STUDI LITERATURE

A. *Industry*

The industry is regarding all human activities in the field of the economy that are productive and commercial. According to Arsyad (2007) industrial development is a function of the main purpose of the welfare of the people is not a independent activities for just physical reached only. The industrial sector is believed as a sector that can lead other sectors in an economy toward advancement.

In his book that in the paper by Dumairy (Dumairy, 2009) industrial products always have the basis of exchange rate (term of trade) that high or more profitable and create added value is greater than the other sector products. In Figure 1 is the importance of creative industries[1].

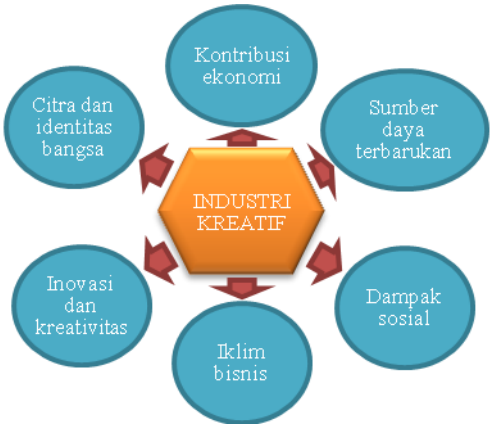

Fig. 1 The Importance of Creative Industries

Creative industry is the industry that is derived from the utilization of creativity, skills and the talent of the individual to create prosperity and jobs with produce and exploit the power of creation and the power of the individual copyright

(p-issn: 2579-5988, e-issn: 2579-597X)

notice[2].

B. Data Mining

Data mining are activities extract or mine knowledge from large amounts of data, this information will be very useful for the development. Data mining aims to find the patterns and rules that is found in the data from the pattern and the rule can be done decision-making and predict the effect the decision[3]. There are three types of the method used to identify patterns in data:

1. Simple model (query SQL based, OLAP, human considerations)
2. The Model (regression, decision trees, clustering)
3. The complex Model (network of nerves, another rule induction).

In Figure 2 is the step of processing Knowledge Discovery in Databases (KDD).

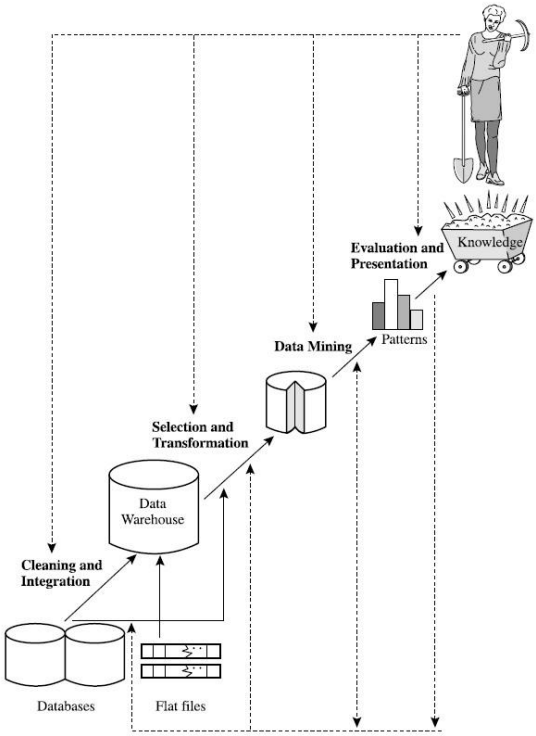

Fig. 2 Step of KDD Process[4]

C. Clustering Method

Clustering is a process of participant grouping the data into classes or cluster the cluster based on a similarity of the attribute - the attribute among groups the data. The goal of the process of clustering is to group the data into a cluster, so that objects in a cluster have a very great similarity with other objects on the same cluster, but are not very similar to the object on the other clusters. One of the characteristics of clustering is good or optimal performance is if the produce cluster that contains data with the level of similarity (similarity) is high on the cluster and the same level of low similarity on different clusters[5].

Cluster analysis is the analysis of the statistics that aims to classify the objects of amatan become some groups based on the variables which are observed. The process of clustering objects based on the characteristics of the similarity between the objects of the object. The object can be in the form of products (goods and services), objects (plants or other) and(respondents or consumers). The object will be classified into one or more cluster (group) so that objects that are located in one cluster will have similarities with one another (Johnson and Wichern, 2002). Cluster analysis can be applied in the field of science,planners of marketing, social and industry[6].

D. K-Medoids

K-Medoids algorithm, also known as coupling around Medoids. K-Medoids algorithm is the coupling method to retrieve the value of the average of the objects in a cluster as a point of reference, medoid screened is the object in a cluster is the most concentrated[7].

Each type of clustering has the advantage and to the respective lemahan. Types of clustering is able to deal with the influence of outlier namely median, so that they develop an alternative method that can group data that contain outlier namely k-medoid.

Outlier according to Johnson & Winchern (1998) is a observations on a series of visible data is not consistent sisaan from the data. On the analysis of the cluster the existence of outlier can cause cluster formed become not representative (Barnet & Lewis 1994).

K-medoid is one of the methods of clustering no hierarchical using the median as the center of clusternya. K-medoid is one of pengclusteran technique that is similar to the k-means. But the difference is adequately constructed on the k-medoid data/objects selected as the center of the cluster (medoid). This method uses the data that is located in the middle of the cluster, then this method is more robust against outlier compared with k-means method (Kaufman & Rousseuw 1990). Medoid can be interpreted as an object of a cluster that has an average of the smallest distance to other objects, in other words that is the object which is located in the middle of the cluster data. This analysis minimize inequality had each object in the cluster using the value of absolute error (E)[8].

$$E = \sum_{c=1}^{k} \quad \sum_{i=1}^{n_c} |p_{ic} - O_c| \qquad (1)$$

Description :
n_c = number of objects in the cluster to the c
p_{ic} = non medoids object i in the cluster to the c
O_c = medoids values in the cluster to the c

E. Preprocessing Data

Before performing the data mining needs to be done pre processing to ensure the data will be processed in data mining is a good data. The Data quality is less good, can be caused by several things namely

1. Incomplete: namely data that lack the attribute value or only contains data aggregate
2. Noisy is data that still contains errors and contains data that is not fair (Anomalies/outliers) because the data collection instruments used may be incorrect, human error or computer that occurred at the time of data entry, error in data transmission.
3. Inconsistent, namely data that contains the difference in code and the name or in short data not consistent[9].

In Figure 3 is the step of preprocessing data before performing data mining are below.

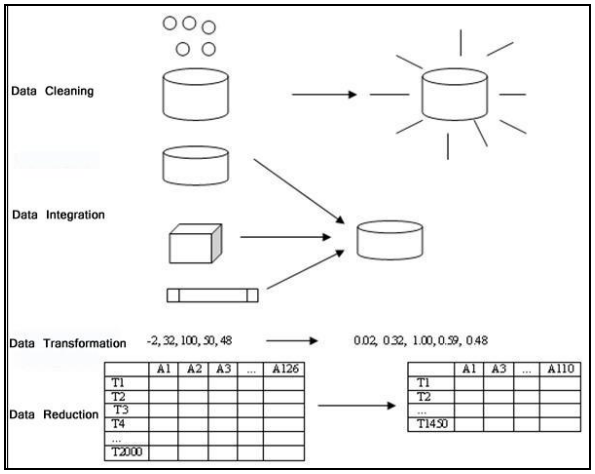

Fig. 3 Preprocessing Data [10]

Preprocessing can be done with some techniques namely

1. Data Cleaning

   The data cleaning will be done among others fill the missing value, identify outlier, handle data noise, corrects data is not consistent and complete data redudansi problem due to data integration.

2. Data Integration

   Data integration is a step to combine data from a number of sources. Data integration is only done if the data comes from different places (source data not only from the 1 place).

3. Data Transformation

   Data Transformation namely change a data so that the obtained data that is more qualified. That will be done among others eliminate noise from the data (smoothing), agregation data, generalisation data, normalisation data, and the establishment of an attribute/features.

4. Data Reduction

   The data reduction namely steps to reduce the dimensions, or attribute the number of data. That will be done among other data aggregation cube,discretisation, dimension reduction and data compression[11].

III. RESEARCH METHODOLOGY

A. Data Source

The Data obtained in this research can be grouped into two types according to the source of the data the primary data and secondary data. The primary data in the form of the data obtained from the research is data Disperindag Bali Province 2012. The following are some of the data used in research.

NO	NAMA PERUSAHAAN	NAMA PRODUK	TENAGA KERJA (ORANG)	NILAI INVESTASI (RP.000)	NILAI PRODUKSI (RP.000)	NILAI BB/BP (RP.000)
1	UD ANANDA	DAGING KALENG	9.00	11,450,000.00	552,000,000.00	280,200,000.00
2	PT SATRIA PANGAN SEJATI	DAGING KALENG	30.00	179,400,000.00	690,000,000.00	402,000,000.00
3	UD SUBAJAYA MITRA	DAGING	8.00	91,750,000.00	1,008,000,000.00	447,000,000.00
4	PT. USIKADA	SOSIS, BAKSO AYAM	10.00	88,100,000.00	191,460,000.00	59,040,000.00
5	UD. SETYA YOGA	PENGOLAHAN DAN PENGAWETAN	6.00	40,000,000.00	2,040,000,000.00	1,855,000.00
6	GARMENT GARAGE	PENGAWETAN DAN PENGOLAHAN DAGING	5.00	216,245,000.00	500,000,000.00	386,139,000.00
7	WEDA PERKASA	PENGOLAHAN DAGING	7.00	18,860,000.00	216,000,000.00	76,290,000.00
8	DUMADAK IDA JAYA	PENGASINAN ATAU PEMANIS BUAH DA SAYURAN	5.00	22,630,000.00	375,000,000.00	233,866,000.00
9	FRESH MAX	PENGOLQHAN DAN PENGAWETAN BUAH DAN SAYURAN	2.00	23,750,000.00	288,600,000.00	151,108,000.00
10	DENDENG SEMARA	PENGOLAHAN DAN PENAWETAN DAGING	2.00	10,740,000.00	64,800,000.00	30,792,000.00
11	CV DUMAK IDA JAYA	SELAI STRAWBERY,MARMALADE	15.00	4,180,000.00	60,840,000.00	17,028,000.00
12	UD NYAMA BALI	MINYAK VCO	3.00	2,000,000.00	114,000,000.00	46,800,000.00
13	CV NYUH GADING	MINYAK VCO	4.00	1,515,000.00	38,400,000.00	12,000,000.00
14	TEGAL SARI	MINYAK GORENG	15.00	2,238,000.00	24,375,000.00	20,000,000.00
15	VOC (MINYAK KELAPA MURNI)	MINYAK NABATI	3.00	633,000.00	14,400,000.00	12,504,000.00
16	GERAI PUDING	MAKANAN DARI COKLATDAN KEMBANG GULA	2.00	2,099,000.00	7,920,000.00	26,016,000.00
17	UD.JAYA ABADI	MAKANAN DARI COKLAT	3.00	5,875,000.00	53,040,000.00	103,392,000.00
18	KHATULISTIWA KOPI DAN KAKAO	MAKANAN DARI COKLAT	3.00	48,493,000.00	210,000,000.00	88,740,000.00
19	UD ALEP	SUSU KEDELAI	5.00	12,510,000.00	61,200,000.00	4,800,000.00

Fig 4. Disperindag Data

B. Research Variables

Variable research has special cirri among others have variable values that vary the variables distinguish one object with other objects and variables must be measured[12]. Some of the variables in this research is as follows.

1. The number of manpower in Bali Province Disperindag data
2. The value of investment in Bali Province Disperindag data
3. The value of the production in Bali Province Disperindag data
4. The value of raw materials in Bali Province Disperindag data
5. The percentage of export data Disperindag Bali Province

C. K-Medoids Algorithm

The following is the algorithm k medoid[13]

1. Given k
2. Randomly pick k instances as initial medoids
3. Assign each instance to the nearest medoid x
4. Calculate the objective function
5. the sum of dissimilarities of all instances to their nearest medoids
6. Randomly select an instance y
7. Swap x by y if the swap reduces the objective function
8. Repeat (3-6) until no change

In Figure 5 is a flowchart of K-Medoid algorithm.

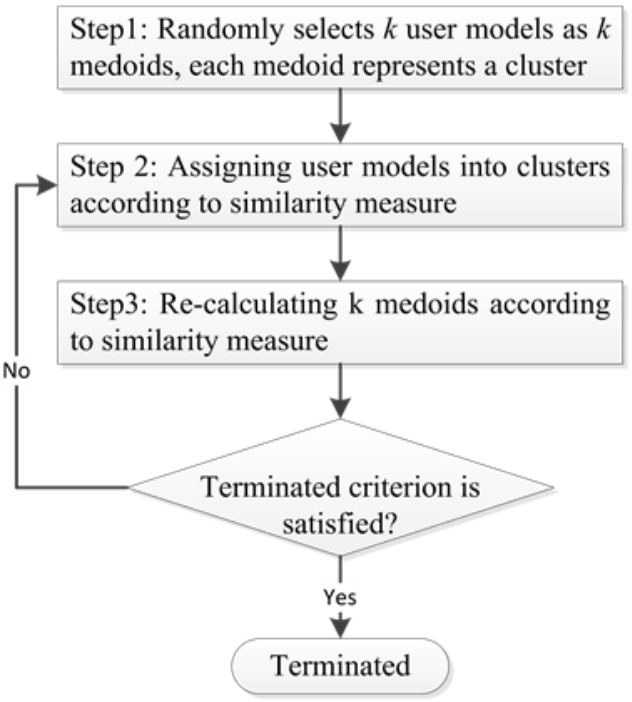

Fig. 5 Flowchart of K-Medoid Algorithm[14]

IV. Results and Discussion

A. Preprocessing Data

Data preprocessing process is done using the tool rapid miner als. The first step is to take the data that made, or data provided by the users can be done with a click the Operator tab select Import →Data→Read Excel. This is used to import the data owned. There are various kinds of choice adjusted with data that has no (in this using Excel). Do Click and drag "Read Excel" to play in the process. Then choose the type of data cleansing. As seen in the figure 6.

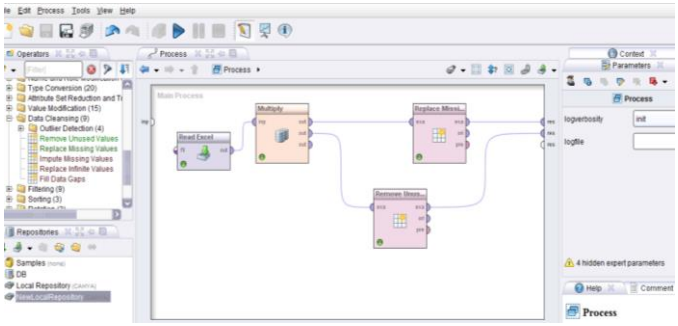

Fig. 6 Preprocessing Data in Rapid Miner

B. Clustering Process

The following is the process of clustering using k-medoid method. In this research done trial clustering with cluster number of as much as 2 and the number of the cluster as much as 3. This figure with the number of as much as 2 clusters. In Figure 7 is an example of clustering process detail in the K-Medoid method.

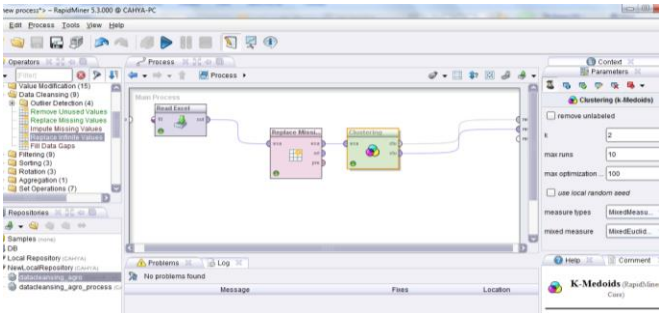

Fig 7. Process Clustering

C. Clustering Results

The following are the results of the process of clustering using k-medoid.

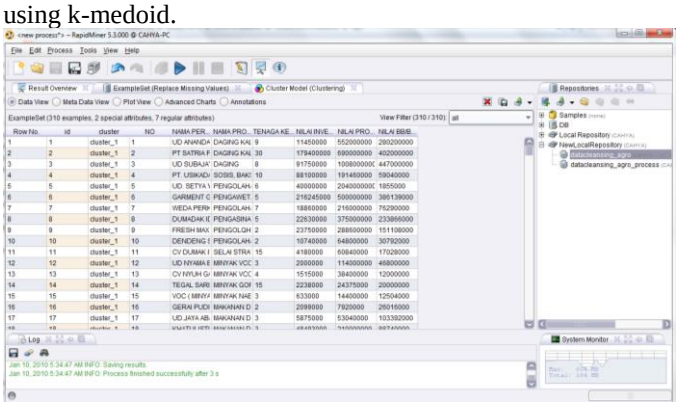

Fig 8. Clustering Result

The results of clustering with the number of cluster 2 produces the first group contains 3 members, the second group contains 307 members. The results of clustering with the number of cluster 3 produces the first group contains 85 members, the second group contains 222 members and the third group numbered 3 members.

D. Cluster Evaluation Method

Purity method are cluster evaluation method used to calculate the purity of a cluster which represented as members of the cluster are the most appropriate (suitable) in a class[15]. The value of purity is closer to the value of 1 indicates the better cluster obtained.

$$purity\ (\Omega, C) = \frac{1}{N}\sum_k max\ |w_k\ \cap\ c_j| \qquad (2)$$

Where $\Omega = \{w_1, w_2, \ldots w_k\}$ is the set of clusters and $C = \{c_1, c_2, \ldots c_j\}$ is the set of classes.

Cluster no	Cluster_0	Cluster_1
1	3	-
2	-	307
Purity = 1		

Purity = (1/310) x (3+307)
= (1/310) x 310
= 1

Cluster no	Cluster_0	Cluster_1	Cluster_2
1	85	-	-
2	-	222	-
3	-	-	3
Purity = 1			

Purity = (1/310) x (85+222+3)
= (1/310) x 310
= 1

The value of the purity of 1 meaning that the quality of the cluster that is formed in this research can be said to have been good and there are no errors in the classification of data.

V. Conclusion and Future Works

The conclusion obtained from the research done is as follows the results of clustering with the number of cluster 2 produces the first group contains 3 members, the second group contains 307 members. The first group are classified as productive because it has a combination of the value of the production of the most high

The results of clustering with the number of cluster 3 produces the first group contains 85 members, the second group contains 222 members and the third group numbered 3 members. The third group are classified as productive because it has a combination of the value of the production of the most high

The results of clustering have the quality of purity worth 1 means good cluster quality.

For future research can combination k-medoid method with another method to produce better cluster. In addition, future research can implement k-medoids method in web application that responsive in mobile application too.

References

[1] H. Hasan, "Analisis Industri Perbankan Syariah Di Indonesia," *J. Din. Ekon. Pembang.*, vol. 1, no. 1, pp. 1–8, 2012.

[2] adha panca Wardanu and M. Anhar, "STRATEGI PENGEMBANGAN AGROINDUSTRI KELAPA SEBAGAI UPAYA PERCEPATAN EKONOMI MASYARAKAT DI KABUPATEN KETAPANG Adha Panca Wardanu 1) dan Muh Anhar 2) 1)," *Agritech*, vol. 1, 2014.

[3] H. A. Edelstein, *Introduction to Data Mining and Knowledge Discovery 3rd edition*, vol. 2. 1999.

[4] U. Fayyad and P. Stolorz, "Data mining and KDD: Promise and challenges," *Futur. Gener. Comput. Syst.*, vol. 13, no. 2–3, pp. 99–115, 1997.

[5] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

[6] P. Berkhin, "A Survey of Clustering Data Mining Techniques," *Group. Multidimens. Data*, no. c, pp. 25–71, 2006.

[7] R. C. Balabantaray, C. Sarma, and M. Jha, "Document Clustering using K-Means and K-Medoids," *arXiv1502.07938 [cs]*, 2015.

[8] A. Bhat, "K-Medoids Clustering Using Partitioning Around Medoids for Performing Face Recognition," *Int. J. Soft Comput. Math. Control*, vol. 3, no. 3, pp. 1–12, 2014.

[9] I. Molloy, N. Li, Y. A. Qi, J. Lobo, and L. Dickens, "Mining roles with noisy data," *Proceeding 15th ACM Symp. Access Control Model. Technol. - SACMAT '10*, p. 45, 2010.

[10] J. Han, M. Kamber, and J. Pei, "3 - Data Preprocessing," in *Data Mining (Third Edition)*, 2012, pp. 83–124.

[11] J. Han, M. Kamber, and J. Pei, "Data Preprocessing," in *Data Mining Concept and Techniques*, 2012, pp. 83–134.

[12] C. Ramayani, "PENGARUH INVESTASI PEMERINTAH, INVESTASI SWASTA, INFLASI, EKSPORT, TENAGA KERJA DAN PRODUKTIVITAS TENAGA KERJA TERHADAP PERTUMBUHAN EKONOMI DI INDONESIA," *J. Econ. Res. Econ. Econ. Educ.*, vol. 1, no. 2, 2015.

[13] P. S. Lai and H. C. Fu, "Variance enhanced K-medoid clustering," *Expert Syst. Appl.*, vol. 38, no. 1, pp. 764–775, 2011.

[14] S. Harikumar and P. V. Surya, "K-Medoid Clustering for Heterogeneous DataSets," in *Procedia Computer Science*, 2015, vol. 70, pp. 226–237.

[15] P.-N. Tan, M. Steinbach, and V. Kumar, "Chap 8 : Cluster Analysis: Basic Concepts and Algorithms," *Introd. to Data Min.*, p. Chapter 8, 2005.