Implementation of the K-Means Clustering Method in Data Grouping Sales In Asia Africa Dentures Dental

Sonibe Halawa¹, Rita Hamdani²

¹²Engineering Informatics Program, STMIK Pelita Nusantara Medan, Jl. Iskandar Muda 1 Medan, North Sumatra 20154, Indonesia

E-mail: sonibehal@gmail.com

Abstract - Data mining can be applied to explore the added value of a set of data in the form of knowledge that had been unknown to them manually. There are several techniques used data mining, one of the data mining techniques is clustering. Clustering can be used for grouping to something. As can group sales data that is most desirable, and others. Examples of companies engaged in the sale is a dental in Africa. Asia Africa Dental is one area of business engaged in the sale of false teeth. Asia Africa Dental these every day to meet the needs of consumers. But Asia Africa Dental lacking in reviewing products sold. What products are needed consumer and data storage is less effective. Thus the need for a system that can support the company in taking decisions quickly and precisely. So in this study, the authors used the application of K-Means Clustering method. To facilitate the author in analyzing the K-Means Clustering The author using the application Weka (Waikato Environment for Knowledge Analysis) .. The result of the calculation Weka (Waikato Environment for Knowledge Analysis) is inserted into the Visual Basic .Net.

Keywords: Data Mining, Clustering, Weka, Visual Basic .Net

1. Introduction

Data mining has some of the techniques used in data mining, one of the data mining technique is clustering. There are two types of clustering methods used in data classification, ie non-hierarchical clustering and hierarchical clustering. K-means clustering as one method of data clustering non-hierarchical partitioning the data into the form of one or more clusters or groups, so that the data which has the same characteristics are grouped into the same cluster and data that have different characteristics are grouped into groups other. Group or cluster acquired a knowledge / information useful for policy users in the decision making process.

Asia Africa Dental is one line of business in the sale of false teeth. Asia Africa Dental these every day to meet the needs of consumers. But Asia Africa Dental lacking in reviewing products sold. What products are needed consumer and data storage is less effective. Thus perluadan a system that can support companies in taking decisions quickly and precisely. In this case analysis was used to overcome the above problems is to use the standard method of K-Means Clustering. To facilitate the author in analyzing the K-Means Clustering The author using the application Weka (Waikato Environment for Knowledge Analysis). Results of the calculations Weka (Waikato Environment for Knowledge Analysis) is inserted into the Visual Basic .Net.

2. Theory

A. Data Mining

By dawn astuti Hermawati (2018: 3), data mining is a process that employs one or more machine learning techniques (machine learning) to analyze and extract knowledge (knowledge) automatically. Data mining is an iterative and interactive process to find a pattern for the new model are valid (perfectly), useful and understandable in a very large database (massive databases)

- Sahih: can be generalized for the future
- New: what is unknown
- Helpful: can be used to perform an action
- Iterative: requires a process that is repeated
- Interactive: require human interaction in the process

B. K-Means Clustering
By Dawn Astuti Herawati (2018: 127), K-Means Clustering using partitional clustering approach. Each cluster is associated with a centroid (center point). Each dot is placed into the cluster with the nearest centroid. The number of clusters, K, must be specified. The algorithm is basically very simple, namely:

1) Select the K point as the initial centroid
2) Repeat
3) The shape of the K cluster by placing all nearest point,
4) Repeat the calculation of the centroid of each cluster,

Until centroid unchanged

C. Weka (Waikato Environment for Knowledge Analysis)

According Slamet Pujiono TIE et al in Journal ISSN: 1411-3201 Vol. 14 No. 2 JUNI 2013, entitled Analysis of Public Satisfaction Using Weka In Achieve Good Governance in Yogyakarta said Weka is a practical machine learning tools package. Weka is an abbreviation of Waikato Environment for Knowledge Analysis, created at the University of Waikato, New Zealand for research, education and a wide range of applications. WEKA is able to solve problems in real-world data mining, in particular the classification of the underlying machine learning approaches. The software is written in Java class hierarchy with object-oriented methods and can run almost all platforms.

3 Discussion And Testing

3.1 Discussion

A. Data analysis

Of the many sales data obtained, taken 9 types of dentures and selling products based on the years 2015-2018. This data is to be sampled k-means algorithm implementation in determining what products are in demand by customers.

| Table 1. Sales report 2015-2018 |
|----------------------------------|
| No | Produk | 2015 | 2016 | 2017 | 2018 |
|----|--------|------|------|------|------|
| 1  | PFM    | 1556 | 2130 | 2064 | 1639 |
| 2  | EMAX   | 375  | 689  | 703  | 611  |
| 3  | ZIRCONIA | 222 | 247 | 333 | 375 |
| 4  | FRS | 318 | 440 | 608 | 617 |
| 5  | VERTEX | 51  | 82  | 20  | 30  |
| 6  | VALPLAST | 442 | 394 | 302 | 162 |
| 7  | ACRYPLAST | 0 | 156 | 55 | 17 |
| 8  | PROTHESA | 314 | 316 | 361 | 400 |
| 9  | FRAME | 230 | 283 | 182 | 223 |

B. Method Analysis K-Means Clustering

| Table 2. Sales data from 2015 until 2018 |
|-----------------------------------------|
| No | Kode Produk | Produk | 2015 | 2016 | 2017 | 2018 |
|----|-------------|--------|------|------|------|------|
| 1  | P001 | PFM | 1556 | 2130 | 2064 | 1639 |
| 2  | E001 | EMAX | 375 | 689 | 703 | 611 |
| 3  | Z001 | ZIRCONIA | 222 | 247 | 333 | 375 |
| 4  | F001 | FRS | 318 | 440 | 608 | 617 |
| 5  | V001 | VERTEX | 51 | 82 | 20 | 30 |
| 6  | Y001 | VALPLAST | 442 | 394 | 302 | 162 |
| 7  | A001 | ACRYPLAST | 0 | 156 | 55 | 17 |
| 8  | P001 | PROTHESA | 314 | 316 | 361 | 400 |
| 9  | P001 | FRAME | 230 | 283 | 182 | 223 |

a. Specify the number of clusters
   Based on existing sales data annually 4. Cluster that will be the initial centroid with 4 clusters
b. Allocate data into groups randomly
   The sales data are allocated based on product codes in the table 2
c. Calculate the cluster center (centroid) using the K-Means for each group

Center cluster random sampling is

| Table 3. Cluster Center |
|-------------------------|
The formula to calculate the cluster center Euclidean distance is as follows:

\[ D(x_i, y_i) = \sqrt{\sum_{m=1}^{n} (x_{im} - y_{im})^2} \]

a. Distance Object A
   1) Center cluster 1
      \[ D = \sqrt{(1556 - 1550)^2 + (2130 - 2118)^2 + (2064 - 2064)^2 + (1639 - 1639)^2} = 0 \]
   2) Center Cluster 2
      \[ D = \sqrt{(1556 - 1550)^2 + (2130 - 2118)^2 + (2064 - 2064)^2 + (1639 - 1639)^2} = 0.039 \]
   3) Center Cluster 3
      \[ D = \sqrt{(1556 - 1550)^2 + (2130 - 2118)^2 + (2064 - 2064)^2 + (1639 - 1639)^2} = 0.073 \]
   4) Center Cluster 4
      \[ D = \sqrt{(1556 - 1550)^2 + (2130 - 2118)^2 + (2064 - 2064)^2 + (1639 - 1639)^2} = 0.094 \]

On the table is done grouping clusters according to the cluster by using the formula

= IF (AND (C4 < D4, C4 < E4, C4 < F4), 1, IF (AND (D4 < C4, D4 < E4, D4 < F4), 2, IF (AND (E4 < C4, E4 < D4, E4 < F4), 3, 4)))

The process of the above calculation is done for each item sold so can result in the calculation of literacy table 1

| Number | C1    | C2    | C3    | C4    | Cluster |
|--------|-------|-------|-------|-------|---------|
| 1      | 0.00  | 3088.83 | 3287.59 | 3087.14 | 1       |
| 2      | 2521.53 | 673.73  | 772.70  | 1177.33 | 2       |
| 3      | 3146.37 | 334.54  | 266.83  | 109.70   | 3       |
| 4      | 2748.30 | 584.05  | 598.98  | 979.44   | 4       |
| 5      | 3057.18 | 644.10  | 428.56  | 0.00     | 4       |
| 6      | 3088.83 | 0.00    | 273.67  | 644.10   | 2       |
| 7      | 5007.11 | 579.25  | 365.20  | 135.11   | 4       |
| 8      | 5244.42 | 287.59  | 228.86  | 654.32   | 3       |
| 9      | 3267.59 | 275.67  | 0.00    | 428.56   | 3       |

Once can literacy first, then the cluster grouping to get a new cluster center value by using means

a. Grouping value of 1
   1) If the cluster = 1, then write said first (2015), otherwise it will be 0, so based on the first number, then the result is 1556
   2) If the cluster = 1, then write said second (2016), otherwise it will be 0, so based on the second number, then the result is 2130
   3) If the cluster = 1, then write said third (2017), otherwise it will be 0, so by the third number, then the result is 2064
   4) If the cluster = 1, then write said fourth (2018), otherwise it will be 0, so by the fourth number, then the result is 1639

b. Grouping cluster 2
   1) If the cluster = 2, then write said first (2015), otherwise it will be 0, so based on the first number, then the result is 0
   2) If the cluster = 1, then write said second (2016), otherwise it will be 0, so based on the second number, then the result is 0
3) If the cluster = 1, then write said third (2017), otherwise it will be 0, so by the third number, then the result is 0
4) If the cluster = 1, then write said fourth (2018), otherwise it will be 0, so by the fourth number, then the result is 0
c. Grouping cluster 3
1) If the cluster = 3, then write said first (2015), otherwise it will be 0, so based on the first number, then the result is 0
2) If the cluster = 3, then write said second (2016), otherwise it will be 0, so based on the second number, then the result is 0
3) If the cluster = 3, then write said third (2017), otherwise it will be 0, so by the third number, then the result is 0
4) If the cluster = 3, then write said fourth (2018), otherwise it will be 0, so by the fourth number, then the result is 0
d. Grouping cluster 4
1) If the cluster = 3, then write said first (2015), otherwise it will be 0, so based on the first number, then the result is 0
2) If clusters = 4, then write said second (2016), otherwise it will be 0, so based on the second number, then the result is 0
3) If clusters = 4, then write said third (2017), otherwise it will be 0, so by the third number, then the result is 0
4) If clusters = 4, then write said fourth (2018), otherwise it will be 0, so by the fourth number, then the result is 0

The average value of the table can be of seeking value based on year respectively. So the results of the cluster grouping grouping clusters to get the new cluster center value can be seen in Table 5.

Table 5.
Cluster Grouping Based On Literacy 1

| Data No | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---------|-----------|-----------|-----------|-----------|
| 1       | 1556      | 2150      | 1698      | 0         |
| 2       | 0         | 0         | 0         | 0         |
| 3       | 0         | 0         | 0         | 0         |
| 4       | 0         | 0         | 0         | 0         |
| 5       | 0         | 0         | 0         | 0         |
| 6       | 0         | 0         | 0         | 0         |
| *       | 0         | 0         | 0         | 0         |
| Total   | 1556      | 2150      | 1698      | 0         |
| Inside data | 1   | 1         | 1         | 1         |
| Data ran | 1556      | 2150      | 1698      | 0         |

d. Allocate each data to the nearest centroid
e. From the table above can be dilhat allocation data values to the nearest centroid to determine the minimum value of each row. Once known values to the nearest centroid, then the new centroid value can be determined by using the average formula at the beginning of the data in the table above using a centroid value closest to the value of the new cluster center

Table 6.
On the Literacy Center New Cluster 1

| New Cluster center ID | 1556 | 2150 | 2064 | 1639 |
|-----------------------|------|------|------|------|
| Percent Cluster center |      |      |      |      |
| Percent Cluster center 1 | 1556 | 2150 | 2064 | 1639 |
| Percent Cluster center 2 | 158.5 | 157.8 | 163.3 |
| Percent Cluster center 3 | 215.7 | 202.2 | 203.3 |
| Percent Cluster center 4 | 31   | 82   | 36   | 20.5 |

f. Back kelangkah 3, if there are data to move the cluster or if the centroid values above the threshold value, or if the value of the objective function used is still above the threshold. In this manual calculation process ends up literacy to four, which can be seen ditabel below.
g. In this calculation ends in literacy to four with can result in below.

Table 7.
New Cluster Center On Literacy

| Cluster Center | C1  | C2  | C3  | C4  |
|----------------|-----|-----|-----|-----|
| Post Cluster   | 1558| 2130| 2064| 1639|
| Post Cluster   | 546.5| 564.5| 655.5| 614.5|
| Post Cluster   | 304.5| 310.8| 294.75| 292.5|
| Post Cluster   | 31  | 82  | 36  | 23.5|

3.2. Testing With Weka Applications

A. Process PenentuanData

At the time of testing by weka the initial step should be prepared is data, then the data untukuraian used Microsoft Excel application with the Depositary ments with extension CSV (Comma Delimited), describing the picture below.

Fig 1. Data Preparation Tests

B. Conducting Tests By Weka

After preparation Weka data according to the application is activated and select open the file, please find the data that was saved in step 1, the display according to the following picture.
At the time of image data selection finished it otomatisfitur provided by weka is active ie classify, cluster, Associate and others. For this stage due to be completed Cluster.

![Fig 3. Testing Results Weka](image)

In this image display output calculation results according to the Euclidean distance formula used in the K-Means method. In the picture below shows how much data there is 9 and literacy No 4. And the result of the same cluster with the results of manual calculations.

![Fig 4. Visualisasi HasilWeka](image)

In this image display the visualization of the results of calculations using Weka. In figure 4 shows the grouping clustering of the data processed.

4 Conclusion

a. Based on research conducted, it can be concluded that the K-Means algorithm bias are used to group sales data dentures in Asia and Africa Dental based on years of sales.
b. By using the K-Means Clustering using weka the application process more effective and results in a rule with a short time.

5 Reference

[1] Asroni, Ronald Adrian, (May 2015), Application Methods K-Means Clustering Students For Academic Values Based Interface With Weka Case Study in Informatics Engineering UMM Magelang, Scientific Journal Semesta Teknika, are identified 11 April 2018, fromhttp://journal.umy.ac.id/index.php/st/article/view/708/858
[2] Deval Gusron, S. Kom, M. Kom. (June 2018). Creating Storage And Data Processing Applications With Vb.Net. Journal KomTekInfo. UPI "YPTK" Padang, be identified 11 April 2018, fromlppm.upiyptk.ac.id> index.php
[3] Paska Marto Hasugian, 2017, Apriori Algorithm Testing With Weka In Formation Asossiation Application Rule. The logic Penusa Journal. 11 April 2018, from [http://E-Journal.pelitanusantara.ac.id/Index.Php/Mantik/Article/View/341](http://E-Journal.pelitanusantara.ac.id/Index.Php/Mantik/Article/View/341)
[4] Rosa. S and M. Salahuddin, 2016, “Software Engineering Structured and Object Oriented”, Informatics. duo
[5] Slamet Pujiono. 2013. Analysis of Public Satisfaction Using Weka In Achieve Good Governance In the city of Yogyakarta. Ties Journal. Be identified 11 April 2018, fromhttps://ojs.amikom.ac.id/index.php/dasi/article/view/162