Implementation of Clustering Unsupervised Learning using K-Means Mapping Techniques

by Ft 05

Submission date: 06-Feb-2023 07:09PM (UTC-0700)
Submission ID: 2008130472
File name: 5_File230206131208230206131208050308037903.pdf (752.87K)
Word count: 3030
Character count: 15303

IOP Conference Series: Materials Science and Engineering

PAPER · OPEN ACCESS

Implementation of clustering unsupervised learning using K-Means mapping techniques

To cite this article: Aries Abbas *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1088** 012004

Implementation of clustering unsupervised learning using K-Means mapping techniques

Aries Abbas[1*], Pungkas Prayitno[2], Florida Butarbutar[1], N Nurkim[1], Denny Prumanto[1], Fathan Mubina Dewadi[3], Nur Hidayati[4], Agus Perdana Windarto[5]

[1] Universitas Krisnadwipayana, Indonesia
[2] Jakarta Technical University of Fisheries, Indonesia
[3] Universitas Buana Perjuangan Karawang, Indonesia
[4] Universitas Bina Sarana Informatika, Indonesia
[5*] STIKOM Tunas Bangsa, Pematangsiantar, Indonesia

*aries@paramount.co.id

Abstract. The manufacturing sector is one of the major contributors to the Indonesian economy. Human work is still needed on the production floor in the manufacturing industry to ensure a smooth operation. This study explores the use of unregulated learning clustering techniques in data mining in the form of clusters of employees in the production industry. The data collection process is carried out through a survey conducted by the Central Statistical Agency (abbreviated as the BPS) with Url: https:/www.bps.go.id in Large Medium Industries and Micro & Small Industries. The statistics used include 24 industrial classifications, with the number of manufacturing employees in the 2017-2019 industry as a percentage. The unregulated technique of learning clustering is k-means. The Large Cluster (E1) and the Low Cluster are the two labels used (E2). The Davies Bouldin Index (DBI) parameter with a dbi value of 0,929 was used to evaluate the cluster (k=2). The findings showed five manufacturing sectors of the high cluster in 5 cluster and 19 manufacturing sectors of the small cluster in 0 cluster. For each cluster the centroid value is 1.67; 1.64; 1.592 (cluster 1/E1) and 0.348; 0.343; 0.3447 (cluster 0/E2), respectively. The research findings will inform the government to improve labour absorption, which will reduce the unemployment rate by substantial numbers in each manufacturing industry.

1. Introduction

The manufacturing industry sector is one of the sectors that plays an important role in the structural transformation process in the Indonesian economy. Because economic development in Indonesia basically aims to improve the welfare of the community. In Indonesia, the manufacturing industry is one of the largest contributing sectors to the Indonesian economy. According to data from the Ministry of Industry of the Republic of Indonesia (2016), from 2012 to 2015 there was an increase in the contribution of the manufacturing sector from 17.99% to 18.18% for GDP. Therefore, the manufacturing industry is seen as a strategic industry to utilize abundant natural resources, considering that Indonesia has a very high population or workforce, so that this manufacturing industry is able to absorb this large workforce. Based on this, the aim of the research is to conduct cluster mapping of the manufacturing industry in Indonesia. The data used is data on the proportion of labor in each

manufacturing industry sector (percentage). This needs to be done considering the role of the manufacturing industry sector is increasingly important in the development of a country's economy, including Indonesia. The growth of industrial output and the creation of added value in output can absorb large amounts of labor which will reduce the unemployment rate. In computer science engineering, one branch of science[1], [2] that can perform mapping in the form of clustering is unsupervised learning data mining [3]–[7].

Data mining is one of the Unsupervised Learning techniques, where no one can know the predicted results [8]–[10]. The results to be shown depend only on the weight compiled value at the beginning of the construction of the method and the classification of weighted items similarly in a given space or area [11]. In other words, data mining is a learning tool that is ideal for identifying or classifying a pattern of several related items that are not the same [12]. One approach is k-means k-medoids, a data mining method that is very common for commercial, academic, or industrial use. Apart from the benefits of k-means, some previous studies used k-means for cluster mapping. Among them was Agus Perdana Windarto (2020) [11] on clustering combined with classifications in the case of the Covid-19 pandemic in Indonesia. This paper proposes a combination of clustering and classification methods. Clustering results can be performed by mapping 9 provinces in the high cluster (C1 = red zone), 3 provinces in the alert cluster (C2 = yellow zone), and 22 provinces in the low cluster. (C3 = green zone). Besides, the research carried out by Sachin Shinde (2014) [13] on the hunt for scientific papers has been carried out. This paper suggests an improvised architecture that uses the k-means algorithm. The results suggest that the k-means algorithm can be used to achieve better clustering with less complexity. On this basis, it is hoped that the K-means approach would be able to solve the mapping of the proportion of employees in the manufacturing sector in Indonesia.

2. Methodology

The k-means approach is the simplest and most commonly used method of dividing datasets into "k" classes [4]. The aim is to divide objects into groups with different characteristics from one category to another [14], [15]. The k-means method is often referred to as the unsupervised modelling technique. An example of cluster results using the k-means approach is as follows:

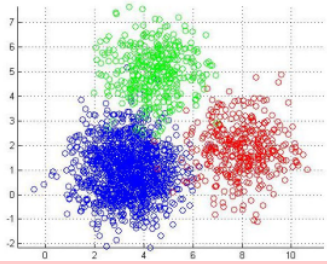

Figure 1. Clustering by using the k-means form

The method used to maximize the usage of k-means is as follows [12]:
Phase 1. Step 1. Identify the number of clusters.
Phase 2. Step 2. Allocate data to clusters randomly.
Phase 3. Step 3. Calculate the centroid/average data in each cluster.
Phase 4. Step 4. Allocate the data to the nearest centroid/average
Phase 5. Step 5. Return to Phase 3, if the data is still moving clusters or if the centroid value changes.

In the implementation of unsupervised learning clustering on the proportion of workers in the Manufacturing Industry Sector with the k-means mapping technique, the data collection process is carried out through a survey conducted by the Central Statistical Agency (abbreviated as the BPS) with Url: https:/www.bps.go.id in Large Medium Industries and Micro & Small Industries. The statistics used include 24 industrial classifications, with the number of manufacturing employees in the

2017-2019 industry as a percentage. In this case the analysis process uses the help of Rapid Miner software. The following raw data and data processed are shown in the table below:

Table 1. The data of the proportion of Labor in the Manufacturing Industry Sector (%)

No	Type of Industry	Proportion of Labor in the Manufacturing Industry Sector (%)		
		2019	2018	2017
1	Food industry	3,75	3,68	3,63
2	Beverage Industry	0,30	0,27	0,28
3	Tobacco Processing Industry	0,34	0,36	0,36
4	Textile industry	1,00	1,11	1,13
5	Apparel Industry	2,09	2,04	1,98
6	Leather Industry, leather goods and footwear	0,70	0,61	0,64
7	Timber industry, goods made of wood and cork (excluding furniture) and wicker articles of bamboo, rattan and the like	1,34	1,37	1,34
8	Paper and paper goods industry	0,22	0,23	0,21
9	Printing and reproduction of the recording media industry	0,27	0,29	0,29
10	Manufacture of products from coal and petroleum refining	0,04	0,05	0,05
11	Chemical industry and chemical products	0,32	0,34	0,35
12	Pharmaceutical industry, chemical medicinal products and traditional medicines	0,13	0,11	0,11
13	Rubber industry, rubber and plastic goods	0,48	0,45	0,44
14	Non-metal minerals industry	1,02	0,99	0,99
15	Basic metal industry	0,20	0,18	0,20
16	Metal goods industry, not machinery and equipment	0,54	0,51	0,44
17	Computer industry, electronic and optical goods	0,14	0,14	0,14
18	Electrical equipment industry	0,17	0,14	0,14
19	Machinery and equipment industry	0,17	0,14	0,13
20	Industry of motor vehicles, trailers and semi trailers	0,19	0,17	0,15
21	Other transportation equipment industry	0,20	0,22	0,23
22	Furniture industry	0,63	0,60	0,57
23	Other processing industries	0,58	0,55	0,54
24	Repair services and installation of machinery and equipment	0,15	0,17	0,17

Source Url: https://www.bps.go.id/indicator/9/1217/1/proporsi-tenaga-kerja-pada-sektor-industri-manufaktur.html

3. Results and Discussion

At this stage the analysis process is carried out using the k-means method. Two cluster label mapping is used, namely the high cluster (E1) and the low cluster (E2). The attribute used is data on the percentage of the proportion of workers in the Manufacturing Industry Sector. The following is a k-means method design using Rapid Miner software as shown in the following figure:

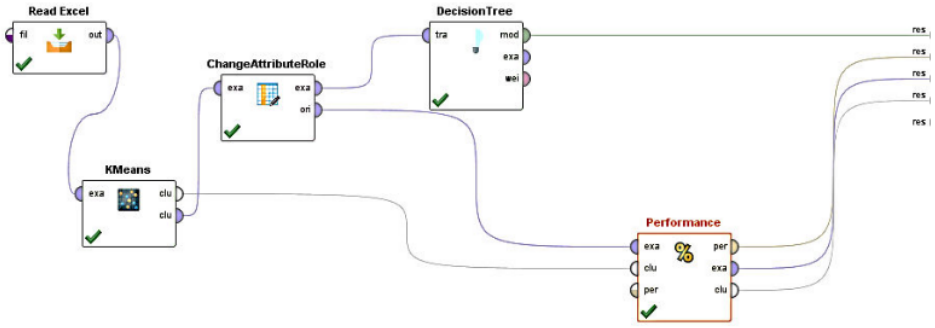

Figure 2. Rapid Miner's model for mapping the proportion of employees in the manufacturing sector

In Figure 2, the data input method uses the read excel tool to enter the data that has been prepared as shown in Table 1. The k-means model is allocated mapping in the form of clusters with input from the previous read excel tool. In addition, performance tools are used to evaluate the intensity of the

clusters that are formed. In this analysis two label clusters were used, namely the high cluster (E1) and the low cluster (E2) for the proportion of employees in the manufacturing industry sector.

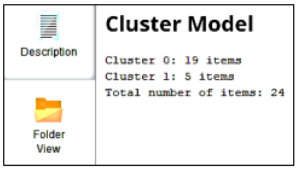

Figure 3. K-means grouping results

Figure 4. The high cluster (cluster_1) and low cluster (cluster_0)

The following are the full clustering results exported from Rapid Miner to Excel as shown in Table 2, where clusters are low (cluster 0) and cluster high (cluster 1).

Table 2. Results of the export the file of Rapid Miner

2019	2018	2017	Type of Industry	label
3.8	3.7	3.6	Food industry	cluster_1
0.3	0.3	0.3	Beverage Industry	cluster_0
0.3	0.4	0.4	Tobacco Processing Industry	cluster_0
1.0	1.1	1.1	Textile industry	cluster_0
2.1	2.0	2.0	Apparel Industry	cluster_1
0.7	0.6	0.6	Leather Industry. leather goods and footwear	cluster_0
1.3	1.4	1.3	Timber industry. goods made of wood and cork (excluding furniture) and wicker articles of bamboo. rattan and the like	cluster_1
0.2	0.2	0.2	Paper and paper goods industry	cluster_0
0.3	0.3	0.3	Printing and reproduction of the recording media industry	cluster_0
0.0	0.1	0.1	Manufacture of products from coal and petroleum refining	cluster_0
0.3	0.3	0.4	Chemical industry and chemical products	cluster_0
0.1	0.1	0.1	Pharmaceutical industry. chemical medicinal products and traditional medicines	cluster_0
0.5	0.5	0.4	Rubber industry. rubber and plastic goods	cluster_0
1.0	1.0	1.0	Non-metal minerals industry	cluster_0
0.2	0.2	0.2	Basic metal industry	cluster_0
0.5	0.5	0.4	Metal goods industry. not machinery and equipment	cluster_1

2019	2018	2017	Type of Industry	label
0.1	0.1	0.1	Computer industry. electronic and optical goods	cluster_0
0.2	0.1	0.1	Electrical equipment industry	cluster_0
0.2	0.1	0.1	Machinery and equipment industry	cluster_0
0.2	0.2	0.2	Industry of motor vehicles. trailers and semi-trailers	cluster_0
0.2	0.2	0.2	Other transportation equipment industry	cluster_0
0.6	0.6	0.6	Furniture industry	cluster_1
0.6	0.6	0.5	Other processing industries	cluster_0
0.2	0.2	0.2	Repair services and installation of machinery and equipment	cluster_0

In Table 2, it can be explained that the results of cluster mapping on the proportion of employees in Indonesia's manufacturing industry market, where the results of the high cluster (E1) are around 21 percent (5 provinces) and the low cluster (E2) is around 79 percent (19 provinces). Here are the last centroid values for the high (cluster 1) and low cluster (cluster 0) values shown below:

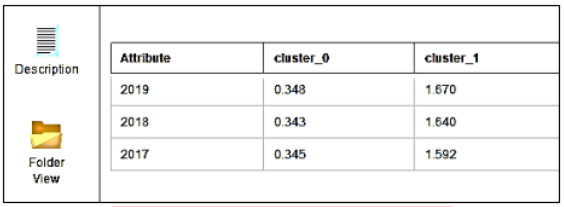

Figure 5. The final centroid results

The following is a mapping picture in the form of scattered plots by mapping on the proportion of employees in Indonesia's manufacturing industry market as shown in the following figure:

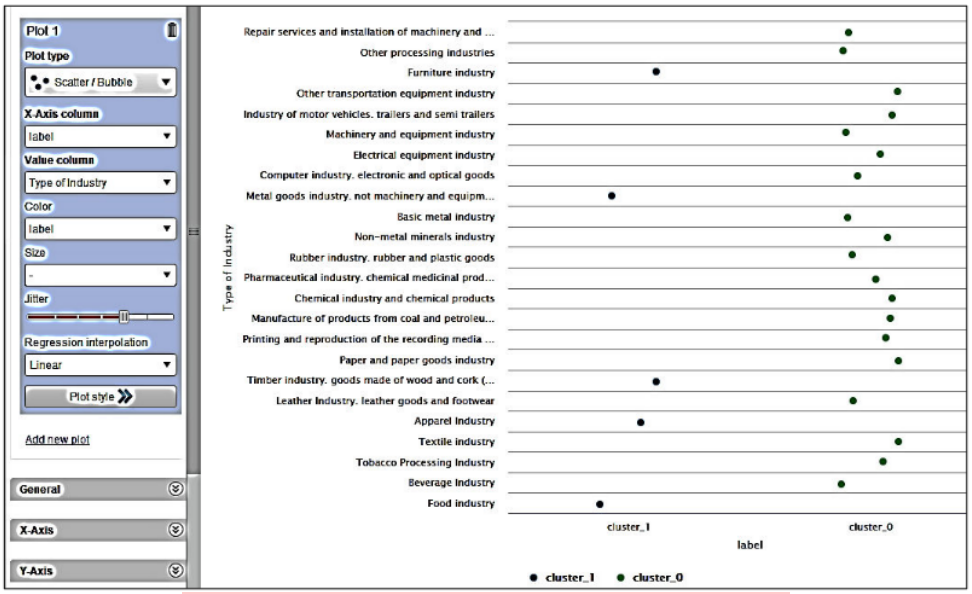

Figure 6. Visualization of clustering results with scatter plotter

In the results of the clustering process, the validation test was used to see the clustering relationship using the Davies-Bouldin method. Tests were carried out on the number of clusters (k=2) with a value of =0.310 as seen in the results of the Rapid Miner picture.

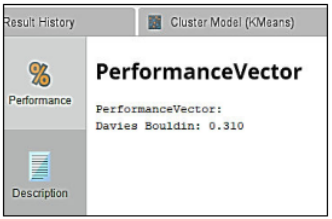

Figure 7. Performance Vector Results

4. Conclusion

Based on the study results, it can be concluded that the implementation of unsupervised learning clustering on the proportion of workers in the Manufacturing Industry Sector using the k-means mapping technique can be applied by using 2 cluster labels (k = 2). This cluster labelling uses the DBI parameter to see the relationship of the cluster formed (DBI = 0.310). Of the 24 manufacturing industry sectors, only about 21 percent have a large workforce absorption. In the future, the absorption of labor in the manufacturing industry sector must be a concern because this can reduce the number of unemployed by creating jobs. Of course, the manufacturing industry is required to create value-added output to absorb the number of workers.

References

[1] W. Rahman, P. T. Nguyen, M. Rusliyadi, E. Laxmi Lydia, and K. Shankar, "Network monitoring tools and techniques uses in the network traffic management system," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2 Special Issue 11, pp. 4182–4188, Sep. 2019.

[2] E. Susanto, Y. Novitasari, W. Rahman, and A. P. O. Amane, "Designing Software to Introduce the Musical Instruments," in *Journal of Physics: Conference Series*, 2019, vol. 1364, no. 1.

[3] F. Rahman, I. I. Ridho, M. Muflih, S. Pratama, M. R. Raharjo, and A. P. Windarto, "Application of Data Mining Technique using K-Medoids in the case of Export of Crude Petroleum Materials to the Destination Country Application of Data Mining Technique using K-Medoids in the case of Export of Crude Petroleum Materials to the Destination C," 2020.

[4] A. P. Windarto *et al.*, "Analysis of the K-Means Algorithm on Clean Water Customers Based on the Province," *J. Phys. Conf. Ser.*, vol. 1255, no. 1, 2019.

[5] D. Hartama, A. Perdana Windarto, and A. Wanto, "The Application of Data Mining in Determining Patterns of Interest of High School Graduates," *J. Phys. Conf. Ser.*, vol. 1339, no. 1, 2019.

[6] A. Waluyo, H. Jatnika, M. R. S. Permatasari, T. Tuslaela, I. Purnamasari, and A. P. Windarto, "Data Mining Optimization uses C4.5 Classification and Particle Swarm Optimization (PSO) in the location selection of Student Boardinghouses," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874, no. 1, pp. 1–9, 2020.

[7] M. Widyastuti, A. G. Fepdiani Simanjuntak, D. Hartama, A. P. Windarto, and A. Wanto, "Classification Model C.45 on Determining the Quality of Custumer Service in Bank BTN Pematangsiantar Branch," *J. Phys. Conf. Ser.*, vol. 1255, no. 012002, pp. 1–6, 2019.

[8] A. P. Windarto, "Penerapan Data Mining Pada Ekspor Buah-Buahan Menurut Negara Tujuan Menggunakan K-Means Clustering," *Techno.COM*, vol. 16, no. 4, pp. 348–357, 2017.

[9] B. Supriyadi, A. P. Windarto, T. Soemartono, and Mungad, "Classification of natural disaster prone areas in Indonesia using K-means," *Int. J. Grid Distrib. Comput.*, vol. 11, no. 8, pp. 87–98, 2018.

[10] Y. Elgimati, "Weighted Bagging in Decision Trees: Data Mining," *JINAV J. Inf. Vis.*, vol. 1, no. 1, pp. 1–14, Oct. 2020.
[11] A. P. Windarto, U. Indriani, M. R. Raharjo, and L. S. Dewi, "Bagian 1: Kombinasi Metode Klastering dan Klasifikasi (Kasus Pandemi Covid-19 di Indonesia)," *J. Media Inform. Budidarma*, vol. 4, no. 3, p. 855, 2020.
[12] N. Kaur, J. K. Sahiwal, N. Kaur, and P.- Punjab, "Efficient K-Means Clustering Algorithm Using Ranking Method," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 1, no. 3, pp. 85–91, 2012.
[13] S. Shinde and B. Tidke, "Improved K-means Algorithm for Searching Research Papers," *Int. J. Comput. Sci. Commun. Networks*, vol. 4, no. 6, pp. 197–202, 2014.
[14] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *International Journal of Advance Research in Computer Science and Management Studies*, 2013. .
[15] W. Zhang, "Computational drug repositioning using big data from genetic studies," *J. Appl. Sci. Eng. Technol. Educ.*, vol. 1, no. 1, pp. 1–3, Jun. 2019.

Implementation of Clustering Unsupervised Learning using K-Means Mapping Techniques

ORIGINALITY REPORT

81%	31%	81%	16%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

80%

★ Aries Abbas, Pungkas Prayitno, Florida Butarbutar, N Nurkim et al. "Implementation of clustering unsupervised learning using K-Means mapping techniques", IOP Conference Series: Materials Science and Engineering, 2021

Publication

Exclude quotes	On	Exclude matches	< 15 words
Exclude bibliography	On		