Application of the clustering algorithm to the small and micro industrial companies for mapping regions with k-medoids

Hendra Jatnika1*, Haris Jamaludin2, Auliya Rahman3, I Ibrahim3, Joseph H Sianipar4, Ari Waluyo5*, Abdul Rahman Maulana Siregar6, Q Qomario7

1Institut Teknologi PLN, Indonesia  
2AMIK Veteran, Purwokerto, Indonesia  
3Universitas Islam Kalimantan Muhammad Arsyad Al Banjari Banjarmasin, Indonesia  
4STFT Surya Nusantara, Pematangsiantar, Indonesia  
5Politeknik Dharma Patria, Kebumen, Indonesia  
6Universitas Pembangunan Panca Budi, Medan, Indonesia  
7STKIP Al Islam Tunas Bangsa, Bandar Lampung, Indonesia

Email: h.jatnika@itpln.ac.id*, auliyaharman@uniska-bjm.ac.id, ibrahim@uniska-bjm.ac.id, ari.hardware@yahoo.com*, abdulrahman@dosen.pancabudi.ac.id, Qomario@stkipalitb.ac.id

Abstract. The largest segment of national economic actors is the micro and small enterprises (MSEs). The research aims to analyze the number of micro and small companies in Indonesia using data mining techniques to map a cluster. The dataset used consists of three-four records of the central Statistical Bureau (SourceUrl: https://www.bps.go.id/), composed of the micro and smaller companies in 2017-2019. The k-medoids method is the solution used in cluster mapping. The average number of small and micro enterprises for 2017-2019 is the attribute used. The numbers of clusters were determined with the Davies Bouldin Index method (DBI), where k = 2 is the best value (0.111). In accordance with the results of the cluster (k=2), the label is divided into two (high cluster (clt1) and low group) (clt2). The results of the calculation of k-medoids show that 90% of Indonesia's area is within the low class. In the high cluster are only Central Java, West Java and East Java. Accuracy, accuracy, reminder and f-measurement parameters are 100% demonstrated in the cluster test results. This mapping can be one of the foundations on which to constantly increase the number of micro and small firms, since the role of MSEs in jobs is extremely important, since it is apparent that MSEs have continued to grow considerably.

Keywords: Datamining, Clustering, K-Medoids, Number of Companies, Micro and Small, Indonesia.

1. Introduction
The biggest segment of national economic actors are micro and small businesses (SMEs). In 2018 MSEs remain the largest player in business and reached 51,26 million corporate units, or 99.99% of the business actors in Indonesia, based on data from the State Ministry of Cooperative and MSEs [1]. As far as labor absorption is concerned, MSE's are able to absorb 97.04% of available manufacturing personnel, 97.04% of which are micro-entities that absorb 89.30% of their workforce, whilst small- and medium-sized enterprises each can absorb 4, 26% and 3.48% of the workforce [2]. MSEs are an important source of jobs in Indonesia, and not only reflect the static situation, namely the number of people in those business groups that exceeds the number of large companies, but can also be viewed under dynamic circumstances, namely the rate of growth every year. That is superior to the big effort
The development of MSEs was the main driver of the real sector, and the impact on national economic growth [4] was direct. According to the Central Statistics Bureau report in Indonesia, micro and small enterprises must be given priority to growth, in order to continue contributing to national economic growth and jobs. The mapping data set can be performed with many computer science techniques [5]–[10]. Where the purpose of mapping is to group the data in a 'region' with the same features and in another 'region' with different characteristics [11]–[15].

Data mining [12] describes a collection of techniques to identify in the collected data unknown patterns. Various familiar techniques of data mining such as clustering, classification, estimation and association exist. The clustering technique is a statistical analysis in order to classify data so that the data in the same group have relatively same features as the data in different groups [16]. K-medoids are one of the classification methods [8], [17]. There are many studies related to the k-medoids method that have been carried out to overcome the problems at hand. For example, the mapping of productive industries is being carried out (I. C. Dewi, B. Y. Gautama, and P. A. Mertasana, 2017) [16]. The findings of this research can offer advice on improved industry development, help industry owners to develop the industry to increase the number of products and product quality, and improve the competitiveness of industry owners in developing their products. Then research on Enhanced K-medoids Semantic Web Clustering Algorithm [18]. This report builds a web model of ontology data collection objects based on the algorithmic analysis and improvement of K-center point selection. The results show that the improved algorithm can significantly enhance the exactness of cluster results in the semantian web by means of experimental assessments. It is hoped that, based on these advantages, the results of this research can analyze, using data mining techniques in the form of clusters, the number of micro and small companies in Indonesia.

2. Methodology

Mapping in the form of clusters in micro and small industrial companies in Indonesia, the dataset is sourced from statistical reports presented by the Central Statistics Agency in 2020 with the topic of the number of micro and small industrial companies in 2017-2019 which consists of 34 records. The attributes used are the number of micro companies (X1) and the number of small companies (X2). The dataset is processed using the help of Microsoft Excel by taking the average value (the last 3 years) of the number of micro and small companies. Processed results are analyzed using the help of Rapid Miner software.

![Figure 1. K-Medoids Clustering Flowchart](image)
The following steps are followed by k-medoids (Fig.1):
  a) Find the number of clusters you want;
  b) Choose k data to initialize centroid or medoid, one centroid for each cluster;
  c) Calculate the distance from the original medoid to the data (using Euclidean distance);
  d) Allow and calculate the cost of data for a cluster of the nearest medoid;
  e) Update each cluster medoid with the category values frequently displayed in each class. Compare the resulting cost value. If the cost value is lower then the medoid should be replaced by the new media value, if it is higher then the medoid should not be changed.

3. Results and Discussion

The process of mapping uses various stages. First, the number of clusters (k) that you are going to create is determined. The Davies Bouldin Index (DBI) method is of course used for this. The lowest DBI value based on the cluster. The analysis using the k-medoids method was conducted after the number of clusters has been determined. The following is a concept using software Rapid Miner as shown in Figure 2.

![Figure 2. The first stage design](image-url)

The following is the comparison of cluster values (k) for each DBI. Where the value of k = 2 has the most optimal value, namely 0.111. Following are the complete results and comparison graphs of each k value as shown in Table 1 and Figure 3 below.

| k   | DBI without Normalize | DBI with Normalize |
|-----|------------------------|--------------------|
| k=2 | 0.223                  | 0.111              |
| k=3 | 0.347                  | 0.174              |
| k=4 | 0.657                  | 0.329              |
| k=5 | 0.558                  | 0.279              |
The results of the first stage are followed by the tests. This is done in order to optimize the cluster formed with the best similarity using the Davies Bouildin Index (DBI). The test parameters used are accuracy, precision, recall and f-measuring. The following is a design used to test the results of the clusters formed as shown in Figure 3 below.

Following are the cluster results using the k-medoids method analysis using $k = 2$; measure type = MixedMeasures and mixed measure = MixedeuclideanDistance.
Predefined parameters and final centroid values for high clusters (729,405 and 101,975 (mic and small) and low clusters (14,242 and 35,99, respectively) are the parameters used (micro and small). Among the cluster results, there are three high-cluster provinces (clt1), namely West Java, Central Java and East Java. The other 90% are in the low cluster (clt2).
Figure 5. The result of accuracy, precision, recall and f-measuring

The results show that the precision value of 100% is obtained from 31 low clusters and 3 high clusters. In the meantime, precision, reminder and measurements also reach 100%. The results of the formed cluster are therefore optimal. This is a graphical cluster of the mapping of Indonesia’s micro and small industries as shown below.

Figure 6. The graphical cluster of the mapping of Indonesia's micro and small industries

4. Conclusion

Based on the results of the research, the mapping of Indonesia’s small and micro industries by means of the k-medoid method is concluded. In order to ensure optimum results of Cluster formation, the number of Cluster is determined by considering the DBI value.

References

[1] K. Weya, W. A. Areros, and T. M. Tumbel, “Analisis Perilaku Usaha Mikro Kecil (UMK) di Kabupaten Tolikara Propinsi Papua,” Productivity, vol. 1, no. 3, pp. 267–273, 2020.

[2] R. Hafni and A. Rozali, “Analisis Usaha Mikro, Kecil, dan Menengah (umkm) terhadap Penyerapan Tenaga Kerja di Indonesia,” Ilmu Ekon. dan Stud. Pembang., vol. 15, no. 2, pp. 77–96, 2017.
[3] A. Zahrah and P. A. Wijaya, “Manfaat Eksistensi Usaha Mikro Kecil Menengah (UMKM) Terhadap Unemployment Rate,” *Society*, vol. 10, no. 2, pp. 110–116, 2019.

[4] Ida Nurnida Relawan, “Peran Kelembagaan Usaha Mikro dan kecil (UMK) terhadap Laju Pertumbuhan Ekonomi (LPE) Jawa Barat,” *MIMBAr*, vol. 30, no. 2, pp. 159–170, 2014.

[5] H. Jatnika, M. Huda, R. R. Amelia, and M. A. Manuhutu, “Analysis of Data Mining in the Group of Water Pollution Areas using the K-Means Method in Indonesia Analysis of Data Mining in the Group of Water Pollution Areas using the K-Means Method in Indonesia,” *J. Phys. Conf. Ser. Pap.*., 2021.

[6] A. Abbas, P. Prayitno, F. Butarbutar, and N. Nurkim, “Implementation of clustering unsupervised learning using K-Means mapping techniques Implementation of clustering unsupervised learning using K-Means mapping techniques,” *IOP Conf. Ser. Mater. Sci. Eng. Pap.*, 2021.

[7] H. Sibyan, W. Suharso, E. Suharto, and M. A. Manuhutu, “Optimization of Unsupervised Learning in Machine Learning,” *J. Phys. Conf. Ser. Pap.*, 2021.

[8] W. Wahidin, M. Mugihartadi, T. Hasanah, B. Aviani, and H. Pratiwi, “Application of data mining techniques using the K-Means Method on Unmet Need of Health Services by Province in Indonesia Application of data mining techniques using the K-Means Method on Unmet Need of Health Services by Province in Indonesia,” *J. Phys. Conf. Ser. Pap.*, 2021.

[9] A. Rahmat, H. Hardi, F. A. Syam, Z. Zamzami, and B. Febriadi, “Utilization of the field of data mining in mapping the area of the Human Development Index (HDI) in Indonesia Utilization of the field of data mining in mapping the area of the Human Development Index (HDI) in Indonesia,” *J. Phys. Conf. Ser. Pap.*, 2021.

[10] 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.

[11] A. P. Windarto, J. Na, and A. Wanto, “Bagian 2: Model Arsitektur Neural Network dengan Kombinasi K- Medoids dan Backpropagation pada kasus Pandemi COVID-19 di Indonesia,” vol. 4, pp. 1175–1180, 2020.

[12] 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.

[13] Z. R. S. Elsi *et al.*, “Utilization of Data Mining Techniques in National Food Security during the Covid-19 Pandemic in Indonesia,” *J. Phys. Conf. Ser.*, vol. 1594, no. 1, 2020.

[14] 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.

[15] 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.

[16] I. C. Dewi, B. Y. Gautama, and P. A. Mertasana, “Analysis of Clustering for Grouping of Productive Industry by K-Medoid Method,” *Int. J. Emerg. Technol.*, vol. 2, no. 1, p. 26, 2017.

[17] P. Arora, Deepali, and S. Varshney, “Analysis of K-Means and K-Medoids Algorithm for Big Data,” *Phys. Procedia*, vol. 78, no. December 2015, pp. 507–512, 2016.

[18] W. T. Ji, Q. J. Guo, and S. Zhong, “The improvement of K-medoids clustering algorithm under semantic web,” *Appl. Mech. Mater.*, vol. 380–384, no. Iecsee, pp. 1286–1289, 2013.