

CLUSTERING MODEL OF LECTURERS PERFORMA IN PUBLICATION USING K-MEANS FOR DECISION SUPPORT DATA

Wiji Lestari¹, Sri Sumarlinda²

^{1),2)} Faculty of Computer Science, University of Duta Bangsa Surakarta, Indonesia
Email Korespondensi : wiji_lestari@udb.ac.id

ABSTRACT

This study aims to produce a clustering model using the K-Means algorithm built to map the performance of lecturers' publications. The method used was research and development which includes the stages of data collecting, data preprocessing, clustering process and cluster analysis. The input data consists of 87 with 8 attributes, namely number of articles in Sinta indexed journal, number of articles in Scopus indexed journal, number of citation in Scopus, H-index in Scopus, number articles in Google Scholar indexed journal, number of citation in Google Scholar, H-index in Google Scholar and H-index10 in Google Scholar. The K-Means algorithm was used with 3 clusters and 100 epoches. The results of clustering were distributed in 3 clusters, namely cluster 1 with 17 members, cluster 2 with 32 members and cluster 3 with 38 members. The results of the cluster analysis with the identification of cluster members and the value of the cluster centroid indicated that cluster 1 is lecturers with relatively high publication performance and cluster 3 shows relatively low publication performance. The data from clustering results can be used for decision support model input data for broader lecturer performance performance.

Keywords: clustering, model, lecturers performa, publication, K-Means,

INTRODUCTION

Process Within the scope of higher education, performance appraisal activities are also applied to assess the performance of lecturers, staff, and students. The basis used in the assessment is based on the tri dhrama of higher education. Law Number 20 of 2003 concerning the National Education System states that universities are obliged to provide education, research, and community service. Lecturer is one of the essential components of an educational system in Lestari and Sumarlinda universities, 2020) (Hastuti et al, 2018). Role, duties and responsibilities of the lecturers are very important in achieving national education goals, namely the intellectual life of the nation, improve the quality of Indonesian human, which includes the quality of faith / piety, noble character, and mastery of science, technology, and art, and to realize Indonesian society that advanced, just, prosperous, and civilized (Sugiharti & Muslim, 2016), (Purnomo et al, 2016), (Purnomo et al, 2017). Quality of lecturers performance can be reflected in the productivity and quality in implementing three responsibilities. The Directorate General for Higher Education has imposed a program of lecturers performance load and the evaluation lecturer tasks running according to the criteria set out in the legislation. (Sugiharti & Muslim, 2016) (Mabu & Muhammad, 2016), (Soriano et al, 2018).

The publication is part of a research assignment for a lecturer. Research is an important task for lecturers besides teaching and implementing community development. Some lecturers consider that research, including publication, is a difficult activities. Publication is a means of disseminating research results or thoughts from researchers or lecturers. Publications can be made in scientific journals and proceedings, both on a national and international scale. The scientific publications are also related to indexing agencies such as Scopus, Web of science (WoS), Copernicus, DOAJ, Google Scholar, and others. Journals indexed in Scopus or WoS are prestigious journals for publication. The performance of lecturers' publications is based on the number of publications, the number of citations and the quality of journals for publication. Journal quality is determined by journal indexing institutions (Sugiyarti & Muslim, 2016), (Nursyahputra et al, 2017).

The clustering is used to analyze different data groupings, similar to classification, but these groupings were not determined before the data mining tool was implemented. The clustering divides data into groups based on what data mining tools found. The principle of clustering is to maximize similarity between class members and minimize similarity between clusters (Lestari, 2014), (Jipkate & Gohokar, 2012). Clustering can be performed on data that has several attributes that are mapped as multidimensional spaces and usually numerical types. The grouping tries to find components from the group naturally, based on the proximity of the data, usually measured by the Euclidean distance. Data grouping usually uses unsupervised learning and statistical data analysis techniques. Models are used to classify the same data into homogeneous groups. It is used to operate large data sets to find hidden patterns and relationships to help make decisions quickly and efficiently (Sugiharti & Muslim, 2016), (Lestari, 2014), (Jipkate & Gohokar, 2012).

The era of industrial revolution 4.0 and Society 5.0 is marked by the development of an information technology paradigm for automation and better performance. Big data, artificial intelligence and internet of things (IoT) are 3 new paradigms of computing that play an important role in this era. Clustering is a part of data mining, clustering is usually used for numerical data without label attributes. The application of clustering in education includes mapping academic institutions, mapping the performance of lecturers and employees, student achievement and others (Sugiharti & Muslim, 2016), (Lestari, 2015), (Sumarlinda & Lestari, 2019), (Shovon & Haque, 2012), (Oyelade et al, 2010), (Widiyaningtyas, 2017), (Nursyahputra et al, 2017).

The clustering method used in this research was the K-Means method. K-Means can also be interpreted as a clustering method which is included in the partitioning approach. The K-Means algorithm is a centroid model. The centroid mode is a model that uses the centroid to create clusters. The centroid is the midpoint of a cluster. Centroid is a value. Centroid is used to calculate the distance of a data object to the centroid. A data object is included in the cluster if it has the shortest distance from the cluster centroid. The K-Means algorithm can be interpreted as a simple learning algorithm to solve a grouping problem that aims to minimize double errors (Oyelade et al., 2010), (Velmurugan, 2012), (Cebeci & Yildiz, 2015). The choice of clustering is because the lecturer publication performance data is less valid if labeled, for example low, medium and high, considering the homogeneous input data with a low publication tendency. It is hoped that the results and clustering analysis can be used as input for the decision support management process of campus institutions, especially in assessing the performance of lecturers' research.

THEORITICAL FRAMEWORKS

Clustering

The clustering is a technique used to identify similar objects or individuals by paying attention to several criteria (Lestari, 2014), (Lestari & Sumarlinda, 2020), (Velmurugan, 2012). Cluster analysis is an analysis to classify similar elements as research objects into distinct and mutually exclusive clusters (Lestari, 2014) (Cebeci & Yildiz, 2015). Another definition is an attempt to find a group of objects that represent the same or almost the same (similar) character between one object and another object in a group and have differences (not similar) with objects in other groups (Lestari, 2014), (Ghosh and Dubey, 2013). Cluster analysis is included in the multivariate statistical analysis with interdependent methods. As an interdependent analysis tool, the purpose of cluster analysis is not to link or differentiate with other samples / variables. Clustering is an analysis tool that is useful for summarizing data. In summarizing this data, it can be done by grouping objects based on the similarity of certain characteristics among the objects to be studied (Lestari, 2014), (Lestari & Sumarlinda, 2020).

The clustering and cluster analysis play an important role in segmenting and mapping data or objects. The application of clustering in various fields with different types of objects such as signals, customers, patients, news, plants, images, and others. Clustering includes data mining techniques with nonparametric methods. In general, there are 2 types of clustering, namely partitioning clustering and hierarchical clustering. Clustering usually uses unsupervised learning methods (Lestari, 2014) (Cebeci & Yildiz, 2015).

Algoritma K-Means

The K-Means algorithm is one of the algorithms with the simplest unsupervised learning method that solves the well-known grouping problems. This algorithm procedure follows a simple and easy way to classify certain data sets through a number of clusters (assume j clusters) remain a priori (Velmurugan, 2012), (Cebeci & Yildiz, 2015), (Ghosh and Dubey, 2013). The main idea is to define centroid j , one for each cluster. These centroids must be placed in a sneaky manner as different ones lead to different results. So, the better option would be to place them as far away from each other as possible. The next step is to take each point that belongs to a particular data set and link it to the nearest center of mass. If no points are done, the first steps are taken and the initial grouping is done. At this point, it is necessary to recalculate the new centroid as the center of the cluster resulting from the previous step. After obtaining this new centroid, new binding must be made between the same data point and the nearest new centroid. A loop has been created. As a result of this looping, one will notice that the k centroids change their location step by step until no more changes are made. In other words, the centroid does not move anymore (Cebeci & Yildiz, 2015), (Ghosh and Dubey, 2013), (Jipkate & Gohokar, 2012), (Shovon & Haque, 2012). Finally, this foundation aims to carry out objective tasks, in this case the quadratic error function. K-Means algorithm equations such as formulas (1) and (2) below.

$$d = \sum_{j=1}^k \sum_{i=1}^n \|x_i^j - c_j\|^2 \quad (1)$$

$$c_j^{new} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij} \quad (2)$$

Where:

d = cumulative distance

x = data point
 c = cluster centroid
 c_j^{new} = new cluster centroid in cluster j
 n_j = number of data in cluster j

The purpose of clustering is grouping data into groups called clusters. Data in one cluster has high similarity, data on different clusters is low. The steps in the K-Means algorithm are (Cebeci & Yildiz, 2015), (Shovon & Haque, 2012) .

- (1). Prepare the dataset
- (2). Determine the number of clusters (j)
- (3). Selecting the initial centroid cluster
- (4). Clustering process to form clusters
- (5). Update the new centroid for each cluster
- (6). Repeating steps (3) to (5) until the centroid cluster value remains (convergent).

The K-Means algorithm is significantly sensitive to the initial randomly selected cluster centers. The k-Means algorithm can be run multiple times to reduce this effect. The K-Means is a simple algorithm that has been adapted to many problem domains and it is a good candidate to work for a randomly generated data points. One of the most popular heuristics for solving the k-Means problem is based on a simple iterative scheme for finding a locally minimal solution [1] [9]. This algorithm is often called the k-Means algorithm (Ghosh and Dubey, 2013), (Jipkate & Gohokar, 2012), (Shovon & Haque, 2012).

METHODS

This study implemented the K-Means algorithm for clustering lecturer performance data in publications. The research stages are shown in Figure 1, below.

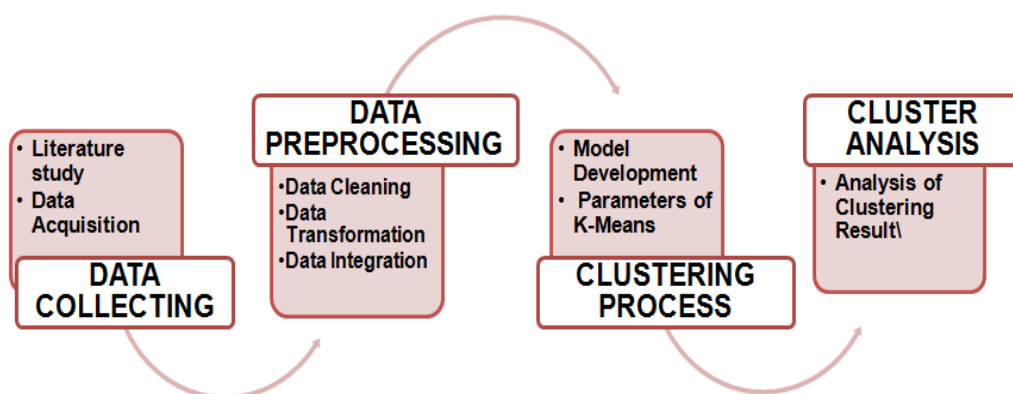


Fig. 1. The Stages of Research

Data collecting was the first step for collecting data on lecturers' publication performance which will be clustered. Input data is taken from indexing institutions with the subject as lecturers at the University of Duta Bangsa Surakarta. The data was accessed for the period of October 2020. The input data amounted to 84 data taken from various journal indexing institutions, especially Google Scholar, Sinta and Scopus. The data attributes for clustering were number of articles in Sinta indexed journal, number of articles in Scopus indexed journal, number of citation in Scopus, H-index in Scopus, number articles in Google Scholar

indexed journal, number of citation in Googe Scholar, H- index in Google Scholar and H-index10 in Google Scholar.

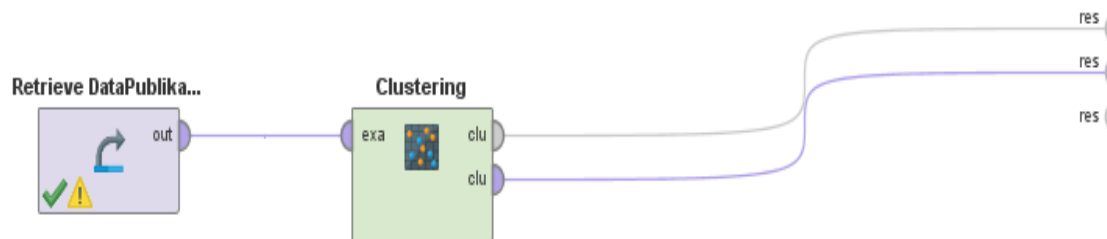
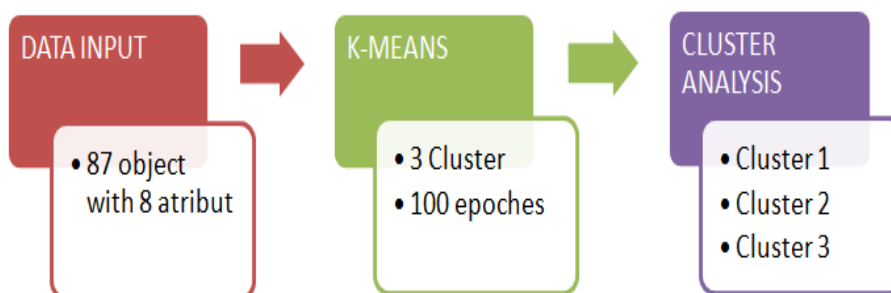


Figure 2. Clustering Using K-Means Model

The preprocessing data consisted of cleaning data and transformation data. Data cleaning was to remove missing data and attributes. Some data was less valid, such as not the number of articles due to articles from other parties being included or articles that have not been indexed in one indexer but have already been indexed in another indexer. Data transformation was changing the data so that it can be used as input data for the clustering process using K-Means.



Fiure.3. Clustering Process

The clustering process used the K-Means algorithm. The clustering parameters used were the number of clusters 3 and 5, while the number of epoches was 100. After the clustering process, the result of clustering analysis was carried out. The clustering results analysis can be used for mapping lecturer performance as well as for decision support data.

DISCUSSION

The clustering model with the K-Means algorithm used 87 objects for input data with 8 numerical variables. The data attributes for clustering were number of articles in Sinta indexed journal, number of articles in Scopus indexed journal, number of citation in Scopus, H-index in Scopus, number articles in Google Scholar indexed journal, number of citation in Googe Scholar, H- index in Google Scholar and H-index10 in Google Scholar. The K-Means algorithm for this clustering process produced 3 clusters with 100 epoches. The clustering

process used Rapidminer. The results of clustering were distributed into 3 clusters, namely cluster 1, cluster 2 and cluster 3. The distribution of the clusters is shown in Figure 4. Below.

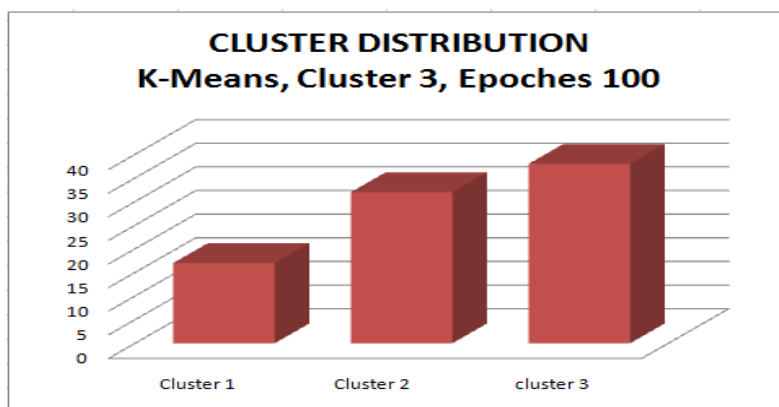


Figure 4. Cluster Distribution K-means: 3 Cluster, 100 epoches

Cluster 1 has 17 members, cluster 2 has 32 members and cluster 3 has 38 members. The membership of each cluster is detailed as in table 1.

Table 1. Number of Cluster Members

No	Cluster	Number of Members
1	Cluster 1	Data no.1, 3, 4, 5, 7, 9, 11, 12, 13, 14, 20, 21, 32, 34, 36, 39, 60. (17)
2	Cluster 2	Data no. 2, 6, 8, 16, 17, 18, 19, 22, 23, 24, 25, 27, 28, 29, 30, 31, 38, 41, 45, 46, 47, 48, 49, 50, 51, 52, 54, 55, 56, 57, 61, 78. (32)
3	Cluster 3	Data no. 10, 15, 26, 33, 35, 37, 40, 42, 43, 44, 53, 58, 59, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 79, 80, 81, 82, 83, 84, 85, 86, 87. (38)

From the cluster analysis, it can be seen that cluster 1 consists of lecturers whose publication performance is relatively high. Cluster 2 medium and cluster 3 consist of lecturers with relatively low performance. Cluster analyst can also look at the centroid cluster per cluster. Cluster 1, the average value of cluster centroid is higher than cluster 2 and cluster 3. The average cluster centroid of cluster 3 is the lowest. The distribution of the centroid cluster is shown in Table 2.

Tabel 2. Centroid of Cluster

Atribut	Cluster 1	Cluster 2	Cluster 3
Number of Documents in Sinta	1.412	1.344	0.263
Number Documents in Scopus	1.118	0.219	0
Number of Citation in Scopus	1.941	0.188	0
H-Index in Scopus	0.529	0.125	0
Number of Documents in Google Scholar	20.059	14.750	5.053
Number of citation in Google scholar	37	14.375	2.553
H-Index in Google Scholar	3	2	0.711
H-Index10 in Google Scholar	0.882	0.344	0.026

The results of the clustering process are expected to be input data for further performance appraisal processes by involving factors related to tridharma, namely teaching, research and community service.

CONCLUSION

The clustering model using the K-Means algorithm was built to map the performance of lecturers' publications. The input data consisted of 87 with 8 attributes, namely number of articles in Sinta indexed journal, number of articles in Scopus indexed journal, number of citation in Scopus, H-index in Scopus, number articles in Google Scholar indexed journal, number of citation in Google Scholar, H-index in Google Scholar and H-index10 in Google Scholar. The K-Means algorithm was used with 3 clusters and 100 epoches. The results of clustering were distributed in 3 clusters, namely cluster 1 with 17 members, cluster 2 with 32 members and cluster 3 with 38 members. The results of the cluster analysis with the identification of cluster members and the value of the cluster centroid indicate that cluster 1 is lecturers with relatively high publication performance and cluster 3 shows relatively low publication performance. The data from clustering results can be used for decision support model input data for broader lecturer performance.

REFERENCES

- Cebeci, Z. and Yildiz, F., 2015, Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures, *Journal of Agricultural Informatics* (ISSN 2061-862X) 2015 Vol. 6, No. 3.
- Ghosh, S. and Dubey, S.K., 2013, Comparative Analysis of K-Means and Fuzzy C-Means Algorithms, ((IJACSA) *International Journal of Advanced Computer Science and Applications*, Vol. 4, No.4.
- Hastuti, I., Purnomo, S. and Lestari, W., 2018, The guidance of Ttechnopreneurship using Expert System Computing Approach Based on Entrepreneurial Values and Multiple Intelligence, *International Journal of Economics, Business and Accounting Research (IJEBAR)*, Vol.2, Issue 3.
- Jipkate, B. and Gohokar, V.V., 2012, A Comparative Analysis of Fuzzy C-Means Clustering and K Means Clustering Algorithms, *International Journal Of Computational Engineering*, Vol. 02, Issue 03.
- Lestari, Wiji. 2014. Sistem Clustering Kecerdasan Majemuk Mahasiswa Menggunakan Algoritma Self Organizing Maps (SOM). *Jurnal Saintech, Jurnal Penelitian dan Pengabdian Masyarakat*, ISSN: 2355-5009, hal : 52 – 58.
- Lestari, W. and Sumarlinda, S., 2020, Implementation of Artificial Neural Network for Clustering of Students Interesting in Computing Fields, *International Journal of Multi Science*, Vol. 1, No, 5.
- Mabu, A.M., Muhammad, F.A. 2016. An Admission Decision Support System for Nigerian Universities. *International Journal of Computer Applications* (0975 - 8887) Volume 133 - No.2, January 2016.
- Nurzahputra, A., Much Aziz Muslim, M.A. and a Khusniati, M., 2017, Penerapan Algoritma K-Means Untuk Clustering Penilaian Dosen Berdasarkan Indeks Kepuasan Mahasiswa, *Techno.COM*, Vol. 16, No. 1.

- Oyelade, O.J., Oladipupo, O.O. and Obagbuwa, I.C., 2010, Application of K-Means Clustering algorithm for prediction of Students' Academic Performance, (*IJCSIS*) *International Journal of Computer Science and Information Security*, Vol. 7, No.1.
- Purnomo, S., Lestari, W., Hastuti, I. 2017. Expert System Design for Guidance of Information Technology (IT) Technopreneurships Based on Students Personal Characteristics. *International Journal of Engineering Research and Technology (IJERT)*, ISSN : 2278-0181.
- Purnomo, S., Lestari, W., Hastuti, I. 2016. The Mapping The Information Technology (IT) Technopreneurships Based of Learning Styles and Multiple Intelligence. *Paper Proceedings International Conference of Education and Technology for Empowering Agents of Change* ISSN : 2541-1578 November 2016. Proceeding of International Seminar , Fakultas Keguruan dan Ilmu Pendidikan, Universitas Kristen Satya Wacana Salatiga, Central java, Indonesia.
- Shovon, M.H.I. and Haque, M., 2012, An Approach of Improving Student's Academic Performance by using K-means Clustering Algorithm and Decision Tree, (*IJACSA*) *International Journal of Advanced Computer Science and Applications*, Vol.3, No. 8.
- Soriano, L.T., Capucan, J.N.B., Palaoag, T.D. 2018. Decision Support Model for Managing Academic Workload in State Universities and Colleges. *International Journal of Information and Education Technology*, Vol. 8, No. 7, July 2018.
- Sugiyarti, E. and Muslim, M.A., 2016, On-line Clustering of Lecturers Performance of Computer Science Department of Semarang State University using K-Means Algorithm, *Journal of Theoretical and Applied Information Technology*, Vol.83. No.1.
- Sumarlinda, S. and Lestari, W., 2020, Design of Decision Support System for Mapping Student Interest in Information Technology and Computing Specialization, *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, Volume 6, Issue 1.
- Velmurugan, T., 2012, Performance Comparison between k-Means and Fuzzy C-Means Algorithms using Arbitrary Data Points, *Wulfenia Journal*, Vol 19, No. 8.
- Widyaningtyas, T., Prabowo, M.I.W. and Pratama, M.A.M, Implementation of K-Means Clustering Method to Distribution of High School Teachers, Proc. EECSI 2017, Yogyakarta, Indonesia, 19-21 September 2017.